GÖTEBORGS UNIVERSITET
HANDELSHÖGSKOLAN

Uncovering Green Innovation: A Topic Modeling Analysis of Climate Change Mitigation Patents

Spring 2025
Graduate School
School of Business, Economics and Law, University of Gothenburg
Institute of Innovation and Entrepreneurship
Author: Oscar Andersson
Supervisor: Daniel Ljungberg

Abstract

This thesis examines the thematic structures and innovation maturity of technologies for mitigating climate change by analyzing the descriptive text of these technologies, specifically patent titles and abstracts. This study proposes a novel methodological approach for mapping and interpreting the development of green technology by combining Latent Dirichlet Allocation (LDA) topic modeling with logistic S-curve analysis. The research adopts an exploratory approach, focusing on the CPC classification method called the Y02 tagging scheme, which categorizes technologies for mitigation or adaptation against climate change. A time analysis of patent applications, modeled with S-curves, demonstrated that general energy storage technologies (Y02E60) entered a maturity phase around 2014, while energy storage using batteries (Y02E60/10) continues to experience accelerated growth. Y02E60 was the second most prevalent CPC group, after Y02E10, which categorizes renewable energy sources. Y02E10 exhibited a persistently stagnant trend in patent applications and was consequently rejected for further analysis.

This study examines 7,211 patents classified under CPC subclass Y02E60/10. LDA disclosed six coherent subjects, including battery cooling systems, chemical composition, safety modulation, and design, underscoring the multifarious technical challenges inherent in battery innovation. This research not only promotes the application of machine learning and innovation theory in patent analysis but also offers practical guidance for R&D strategists and decision-makers in navigating the changing landscape of green technology. In summary, this thesis demonstrates the value of combining text mining and innovation theory to extract useful insights from patent data, contributing to more informed decisions in the rapidly emerging field of climate change mitigation technologies.

Keywords: patent analysis, green innovation, green patents, topic modeling, LDA, S-curve, climate change mitigation, text mining, CPC classification.

Table of Content

List of Figures

List of Tables

List of Equations

1.Introduction

This chapter introduces the urgent global need for climate-focused innovation and positions green patents as a key resource for understanding technological progress. It outlines the limitations of traditional patent analysis methods and motivates the use of advanced text mining techniques like topic modeling. The chapter concludes by framing the research purpose: to explore thematic structures in climate patents

1.1 Background

The necessity to address climate change has reached an urgent juncture. The global agreements known as the Kyoto Protocol (UNFCCC, n.d.a) and the Paris Agreement (UNFCCC, n.d.b) represent the worldwide consensus on the necessity of coordinated efforts to limit global warming. The IPCC's Sixth Assessment Report (2023) underscores the imperative for rapid technological transformation, encompassing the expansion of renewable energy sources and electrification, along with breakthroughs in carbon capture and industrial decarbonization. Despite considerable investments and unprecedented growth in renewable energy capacity, fossil fuels continue to dominate global energy production, and CO_2 emissions persistently increase (Energy Institute, 2024). In order to meet objectives such as the COP28 goal of tripling renewable energy capacity by 2030, it is essential that the global community accelerate innovation across all sectors. Technological advancement, particularly in the domain of green technologies, has emerged as an essential component of climate mitigation and sustainable development strategies. It is essential to comprehend the evolution and maturation of green innovation, as this serves as the foundation for strategic decisions, policy measures, and investment planning across various sectors. As technology evolves, its market potential, need for policy support, and competitive dynamics are affected (Haupt et al., 2007).

Patents play a central role in this understanding. Patents represent one of the oldest institutions in market economies (Kürtössy, 2004), functioning both as legal instruments that grant inventors exclusive rights and as repositories of technical knowledge (Adams, 2019). Their structured and standardized nature makes them valuable for analyzing trends in technological innovation (Hall & Harhoff, 2012). Prior research has utilized several key features offered by patents, including the technical classes assigned by patent offices, the

identity of the inventor, the type and identity of the right holder, and backward and forward citations (Bergeaud et al., 2017). The integration of these characteristics has been demonstrated to facilitate the exploration of the diffusion of knowledge and the interaction between different technological domains (Bergeaud et al., 2017).

1.2 Problem Discussion

The dramatic increase in the number and complexity of patent filings poses challenges for manual patent analysis (Li et al., 2018). The significant increase in the number of patent applications poses substantial challenges for the entire patent system and all users of patent information (Li et al., 2018). Patent classification has a complicated hierarchical structure. Where each patent needs to be assigned one or more labels at the subgroup level. The distribution of patents between categories is very unbalanced, with about 80% of all documents classified in about 20% of the categories (Li et al., 2018). The patent classification of technologies is the responsibility of patent officers, who are experts in their respective fields. However, the applicability of the standardized classification system is constrained by its reliance on the technological state at the time of patent grant, which limits its capacity to anticipate the emergence of novel fields (Bergeaud et al., 2017). Moreover, patent documents are often long and full of technical and legal terms, making it more difficult to analyze them effectively even for domain experts (Li et al., 2018).

The classification system's deficiencies and the limitations of conventional patent metrics have prompted researchers to investigate novel methods for extracting information from textual descriptions such as abstracts and titles of patents. Understanding the content of patent documentation is an important component, which can help to avoid reinventing the wheel and thus save research costs and shorten research time (Li et al., 2018). One of these methods is keyword-based text mining. This method has been employed to assess technical similarity and has been utilized in prior patent analysis research (Yoon, 2008; Kim, Suh, & Park, 2008). According to Park and Yoon (2014), a critical weakness of keyword-based text mining is its reliance on keyword frequency as the sole metric. This approach fails to adequately capture specific technical results and structural relationships between components. To address these limitations, this study employs Latent Dirichlet Allocation (LDA), a statistical topic modeling method that infers thematic patterns from large text corpora (Blei et al., 2003). LDA treats each patent document as a probabilistic mixture of topics, thereby enabling scalable, unsupervised identification of thematic clusters (Blei et al., 2003). This approach

aligns with recent developments in patent analysis, where semantic modeling has proven effective for understanding technological evolution (Gan & Qi, 2021; Wang et al., 2014).

In addition, logistic S-curve modeling is used to evaluate the maturity and identify growth technologies. S-curves have long been established as indicators of technological life cycles and innovation saturation (Haupt et al., 2007; Gao et al., 2013). By tracking topic frequencies over time, the model reveals whether a technology is emerging, growing, or plateauing. In this study, the S-curve is employed to identify patent classifications characterized by growth, with the objective of studying specific terms for a relevant technological domain.

1.3 Purpose

The purpose of this thesis is to examine thematic structures and innovation trends in climate change mitigation technologies using patent data. By combining topic modeling with maturity assessment, the study aims to generate insights relevant to policymakers, researchers, and industry stakeholders. The methodology involves the identification of trends in technological domains related to green technology. A study of patents in the domain of green technology can yield valuable insights for corporate entities and policymakers. The analysis of patent documents has been demonstrated to offer insights into technology trends and the research areas of core technologies (Wang et al., 2014). Companies usually use patents as an effective way to protect their intellectual property rights and the market dominance of new products (Li et al., 2018). Patent data constitutes a significant repository of competitive intelligence, which can be leveraged by organizations to attain strategic advantages. Segmentation of technical subjects is a critical component of competitive analysis, which entails the division of a research domain into numerous sub-fields, with each sub-field comprising multiple technical terms (Wang et al., 2014). Patents function as strategic resources for the management of information and knowledge, with an increasing number of firms leveraging the extensive information available in patent databases to enhance their research and development (R&D) initiatives (Li et al., 2018). These initiatives encompass a wide range of activities, including new product development (Li et al., 2013), technology transfer (Lemley & Feldman, 2016), technological innovation (Lee et al., 2012), technology forecasting (Altuntas et al., 2015) and mergers and acquisitions (M&A) analysis (Park et al., 2013, 2017).

1.4 Research Question

This study is guided by the overarching research question: *What are the dominant thematic structures in climate change mitigation patents?* To address this issue, the present thesis explores two fundamental sub-questions. First, it investigates how unsupervised machine learning can be used to uncover trends in green technological innovation by identifying latent themes in patent texts. Secondly, the study explores the potential of clustering techniques and visualization tools to facilitate the interpretation and validation of the extracted topics, thereby enhancing the comprehension of technological trajectories within the green patent landscape. Consequently, these research questions serve as the analytical foundation of the study and inform the methodological decisions.

Research Question: *What are the dominant thematic structures in climate change mitigation patents?*

- *How can unsupervised topic modeling reveal trends in green technology?*
- *How can clustering and visualization methods help interpret and validate the extracted topics?*

2. Literature Review

This chapter critically reviews the main literature on patents, their classification, measurement, and limitations, specializing in green patents as indicators of sustainable innovation. Drawing on foundational and contemporary sources, the review situates patents within the broader intellectual property framework while questioning assumptions and the effectiveness of patent data as indicators of innovation. Particular attention is paid to green patents and their relevance for addressing environmental challenges. The chapter also discusses new text-based and alternative measures that complement traditional measures and provide a multidimensional picture of technological development and diffusion. The literature is problematized in terms of methodological limitations, conceptual assumptions, and practical implications, laying the ground for further empirical investigations of green technology trends and innovation maturity.

2.1 Patents

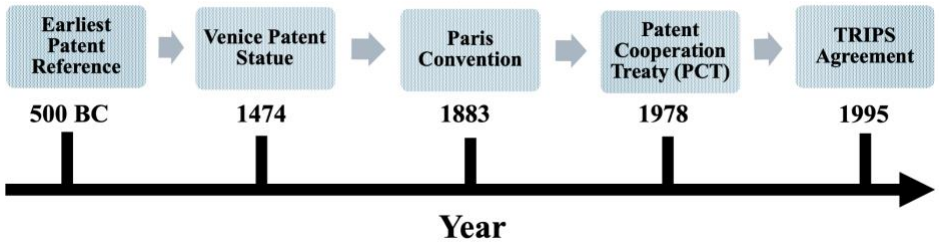
Patents are a cornerstone of intellectual property systems, offering inventors exclusive rights and incentivizing innovation. Their historical evolution—from ancient monopolies to modern international treaties—has shaped today's global IP landscape. Despite their strategic importance, patent systems are contingent on compliance, legal frameworks, and enforcement.

Patents have a long history, with some sources attributing the earliest known references to 500 BC, when it was alleged that the Greeks and Romans granted exclusive rights to individual products (Adams, 2019). The more reliable start of patents was established with the creation of the Venice Patent Statute in 1474, which is regarded as the first modern patent law (Adams, 2019). Since then, significant advancements have been made. A notable turning point was marked by the Paris Convention of 1883, which established the first international treaty, involving eleven nations, that guaranteed the international protection of inventions and established priority rights. This convention laid the foundation for the present-day management of international intellectual property rights by the World Intellectual Property Organization (WIPO) (Adams, 2019). Approximately a century later, in 1978, the Patent Cooperation Treaty (PCT) was established with the objective of reducing costs by simplifying the search process (Hall and Harhoff, 2012). However, the efficacy of laws and regulations is contingent upon their compliance by signatory countries. Consequently, the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) was developed. TRIPS is an international agreement administered by the World Trade Organization (WTO) that promotes compliance with minimum standards in the regulation of intellectual property (Adams, 2019).

Patents are subject to regulation under the category of Intellectual Property Rights (IPR). The World Intellectual Property Organization (WIPO, 2025) defines intellectual property (IP) as “creations of the mind, including inventions, literary and artistic works, designs, symbols, names, and images utilized in commerce.” The provision of protection by patents, copyrights, and trademarks is intended to balance societal benefits by stimulating innovation and rewarding inventors (Hall et al., 2000). A patent grants the inventor the exclusive right to exclude others from exploiting their invention for a predetermined period of time. For a patent to be granted, it must meet the following criteria (Hall et al., 2000): firstly, the invention must be novel within the legal definition of the term, that is to say, it must not have been previously disclosed or published. Secondly, the invention must be non-obvious, and the relevant field experts would not have been able to conceive of it independently.

Furthermore, the invention must be useful; that is to say, it must have potential commercial value.

Figure 1: *Timeline of Important Events in Patent History.*



Note: *This figure presents a chronological overview of key historical events that have shaped the modern patent system. Starting with the earliest known patent reference from 500 BC, the timeline highlights pivotal legal and institutional developments, including the 1474 Venetian Patent Statute (considered the first formal patent law), the 1883 Paris Convention (which introduced international cooperation on industrial property), the 1978 Patent Cooperation Treaty (PCT), and the 1995 TRIPS Agreement under the World Trade Organization, which harmonized intellectual property standards globally.*

2.1.1 Patent Classification

Patent classification systems such as IPC and CPC are critical for organizing and analyzing technological information. Although classification has improved with globalization, the hierarchical structure of systems like IPC introduces technical challenges and inefficiencies, particularly under growing patent application volumes.

Upon completion of a patent application and its subsequent readiness for publication, the next step is the assignment of an appropriate classification to the patent. The classification of patents is a primary function of patent offices worldwide. For illustrative purposes, please refer to Table 1, which provides a selection of patent office examples.

Table 1: *Selection of Patent Offices.*

Patent Offices	Acronym	Jurisdiction
United Kingdom Intellectual Property Office	UKIPO	United Kingdom
European Patent Office	EPO	European Patent Convention (EPC) Contracting States
World Intellectual Property Organization	WIPO	International (PCT member states)
United States Patent and Trademark Office	USPTO	United States
Japan Patent Office	JPO	Japan

China National Intellectual Property Administration	CNIPA	China
German Patent and Trade Mark Office	DPMA	Germany

Note: *This table lists selected national and international patent offices responsible for granting and managing patent rights across jurisdictions.*

Prior to the advent of globalization, patent offices employed their own unique classification system of patents. However, as nations increasingly adopt international classifications, a marked reduction in complexity and associated costs has been observed. The following table offers a concise overview of the various classifications employed, along with the respective patent offices that oversee their administration.

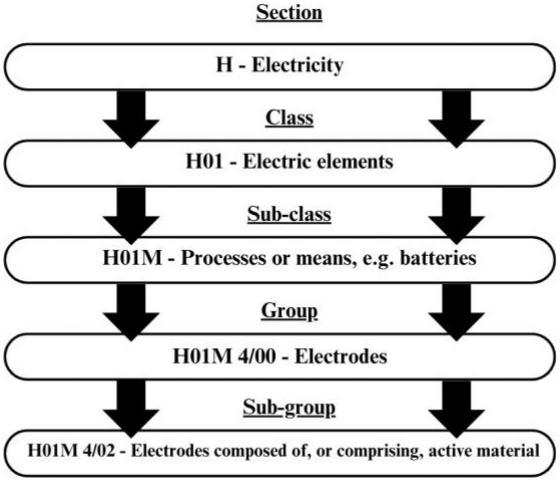
Table 2: *Selection of Patent Classification Systems.*

Classification System	Acronym	Used By	Description
International Patent Classification	**IPC**	WIPO, EPO, UKIPO, USPTO, JPO, CNIPA, etc.	A global classification system administered by WIPO is used to standardize patent categorization worldwide.
Cooperative Patent Classification	**CPC**	EPO, USPTO	A system jointly developed by the EPO and USPTO provides a more detailed classification than IPC.
United States Patent Classification	**USPC**	USPTO	A legacy system formerly used in the U.S., is now replaced by CPC but still referenced in older patents.
Japanese Patent Classification	**F-term/JPC**	JPO	A system used in Japan, with F-terms providing a more refined classification based on functionality.
Chinese Patent Classification	**CPCN**	CNIPA	China's own classification system, although CNIPA also uses IPC and CPC.
German Patent Classification	**DPK**	DPMA	A national system formerly used in Germany, is now largely replaced by IPC.

Note: *The table summarizes common patent classification systems and their administering bodies. While IPC is widely used globally, CPC offers more granular categorization, particularly useful in emerging fields like green technology.*

The most widely adopted classification system is the International Patent Classification (IPC), which has been adopted by over 100 countries and is available in 10 languages (Li et al., 2018). As of now, the IPC classification contains 8 sections, 130 classes, 640 subclasses, 7,400 main groups, and approximately 72,000 subgroups, but the system is updated regularly when necessary. Sections are denoted by capital letters from A to H, encompassing: (A) "Human necessities"; (B) "Performing operations; Transportation"; (C) "Chemistry; Metallurgy"; (D) "Textiles; Paper"; (E) "Solid constructions"; (F) "Mechanical engineering; Lighting; Heating; Weapons; Blasting"; (G) "Physics"; and (H) "Electricity." The second level of the IPC classification is a class represented by a number. At the next level, there are sub-classes, groups, and sub-groups (see Figure 2 for an example of an IPC code).

Figure 2: *Hierarchical Structure of the IPC System.*



Note: *This figure illustrates the hierarchical organization of the International Patent Classification (IPC) system using an example from the domain of battery technology. The classification progresses from broader categories (Section: H – Electricity) to more specific sub-domains (Sub-group: H01M 4/02 – Electrodes composed of, or comprising, active material). The IPC system enables the systematic organization of patent documents based on technical content, facilitating search, comparison, and analysis of innovations across jurisdictions and time periods.*

The Cooperative Patent Classification (CPC) system is a patent classification system developed and maintained jointly by the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO). The CPC draws upon the structure of the IPC system, yet it incorporates a substantially higher degree of detail to enhance the precision of patent categorization and information retrieval (EPO & USPTO, n.d.). The CPC's hierarchical structure mirrors that of the IPC (see Figure 2 for example), consisting of sections (labeled A to H and Y), classes (two-digit numbers indicating specific technical fields), subclasses (a capital letter appended to the class code to indicate narrower areas), main groups (numeric code for main divisions within subclasses), and subgroups (additional digits appended to the main groups to specify, e.g., H01R 12/73) (EPO & USPTO, n.d.). The "Y" section is a distinctive feature of the CPC, utilized for the classification of novel technologies or interdisciplinary domains. Illustrative examples include technologies for mitigating climate change (Y02) and smart grids (Y04S). The CPC has been developed to facilitate efficient searching and rapid classification. The system is subject to continuous updates by the EPO and USPTO in order to reflect technological developments, and all relevant documents are reclassified as necessary (EPO & USPTO, n.d.).

In 2023, a new record in patent applications was set at 3.55 million applications, representing an increase of 2.7% from 2022 (WIPO, 2024). This high volume of applications exerts significant pressure on patent offices and their classification activities, presenting several

challenges. Li et al. (2018) contend that one of the primary challenges lies in the complex nature of the IPC classification system, which is characterized by its intricate, hierarchical structure. The precise allocation of patents to specific subgroups within the IPC system is a time-consuming and technically demanding process, necessitating the expertise of skilled professionals. Furthermore, the distribution of patents across the available categories is marked by significant imbalances, with approximately 80% of all patent documents concentrated in a mere 20% of the available classification categories (Li et al., 2018). This imbalance has the potential to result in inefficiencies regarding the retrieval and analysis of patents. Another challenge is the nature of patent documents, which are often extensive and contain complex technical and legal terminology, making them challenging for even domain experts to analyze efficiently (Li et al., 2018).

2.1.2 Patent Measurements

Patent impact and novelty are traditionally measured via backward and forward citations and patent family size, but these metrics have limitations, particularly in identifying radical innovations. New approaches using text-based measures and machine learning offer more nuanced insights into innovation trajectories and thematic shifts.

2.1.2.1 Traditional Measures

In research centered on the utilization of patent information, a range of indicators have been employed to assess the nature, novelty, and impact of technological inventions. Backward citations refer to earlier patents that a new patent application explicitly cites. These citations serve to indicate the prior inventions or technologies upon which the new patent builds. As asserted by Arts et al. (2013), backward citations can serve as a metric for evaluating the originality of a patent in terms of its nature. A high number of backward citations often suggests that the invention represents an incremental improvement, relying on existing knowledge. Conversely, the absence of backward citations may be indicative of a more original or even disruptive invention, as it does not directly draw from previous patented technologies (Fleming et al., 2007; Arts et al., 2013). Moreover, patents with new combinations of patent classifications may also indicate more creative inventions.

Conversely, forward citations occur when a patent is referenced by later patent applications. These metrics are widely regarded as a measure of the patent's technological relevance and its influence over time. When a later invention cites an earlier patent, it is indicative of the earlier patent's contribution of knowledge or technical features that were either

useful or foundational to the subsequent development. This type of citation functions similarly to how academic papers cite prior research: the more often a patent is cited by future patents, the more it is considered to have shaped ongoing innovation within its technological domain. Consequently, forward citations serve as an indicator of a patent's technological impact, with higher citation counts generally corresponding to greater significance and value (Arts et al., 2013; Trajtenberg et al., 1997). A notable finding is the uneven distribution of citation frequency. Most patents receive minimal to no attention after publication, while a small subset, typically the most innovative or foundational, are cited frequently and influence a wide range of subsequent inventions. These patents, which are frequently cited in academic and industry publications, are often regarded as significant technological advancements and tend to be concentrated within the top 1–5% of the distribution (Arts et al., 2013; Gambardella et al., 2008).

Fischer and Leidinger's (2014) study supported the hypothesis that the number of forward citations could be used to measure the technological quality of the patent. The authors also demonstrated that family size could be used as a measure of the economic value of the patent. Furthermore, the magnitude of the patent family is indicative of the applicant's endeavors to safeguard the patent in multiple jurisdictions. The financial burden associated with patent filing, including translation expenses and additional filing fees to protect the technology, underscores the urgency to ensure its economic value.

Consequently, the most effective measure of the value or usefulness of patents is through citations, yet they do not effectively capture the degree of novelty (Kaplan and Vakili, 2015).

2.1.2.2 Text-Based Measures

In order to measure or identify the novelty of patents, it is necessary to employ methods that differ from traditional indicators such as citations. Kuhn (1970) argued that scientific ideas are found in vocabularies and that changes in ideas can therefore be detected in language changes. Patent citations are intended to capture prior art; however, they do not accurately reflect the technical content of the prior art, thereby hindering the ability to accurately measure the novelty of the technical content (Arts et al., 2021). Furthermore, citations are often an incomplete and biased representation of prior art (Arts et al., 2021). Therefore, to understand the emergence of pioneering new ideas, it is necessary to employ methods that consider the language that represents the innovations. Natural language processing (NLP) and machine learning techniques are now being employed to exploit the extensive documentation and

technical content of patent documents. These techniques have been shown to facilitate the detection of novelty in textual documents, the identification of new technologies, and the tracking of the diffusion and impact of new technologies in patent texts (Arts et al. 2021).

In their study, Kaplan and Vakili (2015) employed "topic modeling," a method that detects latent topics in a collection of patents and identifies the composition of these topics that is optimal for each patent. This approach enables the mapping of different topics, following the creation of new topics— which are then regarded as novel ideas or technological breakthroughs (Kaplan and Vakili, 2015). Latent Dirichlet Allocation (LDA) is a generative probabilistic model that represents documents as mixtures of topics, where each topic is a distribution over words (Blei et al., 2003). The method is useful because of its ability to detect latent thematic structures in unstructured text data. This framework is particularly useful when analyzing patent texts, which often contain high-dimensional and domain-specific language. By extracting thematic information without labeled data, LDA facilitates exploratory analyses of innovation trends and R&D trajectories (Zhang et al., 2021; Pan & Xu, 2023). Through iterative algorithms, LDA can identify topics and word distribution within documents for application in text topic analysis, text classification, and information retrieval (Pan & Xu, 2023).

Previous studies have integrated LDA into large-scale patent analysis to assess technology development and identify emerging areas. For instance, Zhang et al. (2021) used LDA to evaluate blockchain-related patents and investigate how topics and keywords changed through the technology lifecycle stages. The method has also been used in areas such as telecommunications and renewable energy to support strategic foresight and policy recommendations (Pan & Xu, 2023).

Despite this, Arts et al. (2021) underscore significant limitations of LDA in the context of innovation studies. The authors emphasize that LDA bases topic modeling on abstract probabilistic associations between words, which do not necessarily reflect the technical novelty or impact of an invention. The topics generated frequently group words that co-occur, though this does not necessarily capture the functional or inventive content of the patent. Instead, Arts et al. (2021) propose a more direct, text-based method based on natural language processing (NLP) that uses keywords, bigrams, trigrams, and new combinations in patent titles, abstracts, and patent claims, and then tracks how these elements are reused in subsequent patents. In contrast to LDA, this approach captures the actual occurrence and propagation of technical concepts, thereby providing enhanced discrimination capability for identifying groundbreaking patents. Nonetheless, given that the technical domain to be studied was not predetermined, the absence of keywords was a critical component that favored LDA in this study.

2.1.3 Patent Limitations

Despite being widely used as innovation indicators, patents often fail to reflect meaningful impact due to low commercialization rates and strategic decisions to opt for trade secrets. Theoretical foundations and accessibility keep patents central in innovation metrics, but their value is uneven and context-dependent.

Patents are closely associated with innovation, conjuring images of renowned inventors who have historically sought to patent their creations and thereby transform the world or the future within an industry. However, the reality is that only a small percentage of patents affect meaningful change, and a significant majority never progress beyond the conceptual stage (Allison et al., 2003; Sichelman, 2009; Krant, 2023). Furthermore, not all inventions that meet the criteria for patenting ultimately receive such protection. In certain instances, firms may opt to prioritize the maintenance of trade secrets (Mansfield, 1986; Hall et al., 2000). This decision may be influenced by the rapid rate of technological advancement, which can render an invention obsolete before the completion of the patenting process. Additionally, the enforcement of patent rights can be challenging in certain regions, thereby diminishing the efficacy of protection (Mansfield, 1986). In industries where innovation is costly and difficult to replicate, firms may not perceive patents as a necessary protection measure. It is also noteworthy that specific sectors, notably the service sector, are not eligible for patenting or are only eligible to a very limited extent. It must be acknowledged that patent information is indicative of technological developments in sectors where the patenting process is an integral component of the developmental process.

Although patents are frequently regarded as tangible proxies of innovation, their effectiveness as predictors of technological advancement remains a subject of debate. This is due to the disparity between the number of patents filed and those that yield meaningful impact (Allison et al., 2003; Sichelman, 2009; Krant, 2023). This discrepancy raises fundamental questions about the philosophical foundations of intellectual property rights. Hughes (1988) explores these foundations through *Locke's labor theory* and *Hegel's personality theory*, both of which provide insight into why patents continue to be used as innovation indicators despite their limitations. From a *Lockean* perspective, patents serve as a means of rewarding individuals for their intellectual labor, legitimizing their claims to ownership. However, as Fischer and Leidinger (2014) emphasize, the value of patents varies, and metrics such as

forward citations and family size are employed to differentiate patents with greater technological or economic significance. Similarly, *Hegel's personality theory* posits that intellectual property embodies personal expression; nevertheless, the commodification of patents in innovation metrics frequently overlooks this dimension (Hughes, 1988). The persistent reliance on patents as indicators of innovation, despite their varied impact, is indicative of two factors. Firstly, patents' accessibility and structured documentation (Benson & Magee, 2015). Secondly, the enduring belief that intellectual effort and proprietary claims merit protection and recognition within the broader framework of economic and technological progress.

Kim et al. (2012) challenge the conventional wisdom that stronger intellectual property rights universally drive economic growth. Their research emphasizes the importance of aligning IPR policies with a country's technological capabilities. While patents are crucial for promoting innovation in developed countries, utility models (used to protect minor inventions) and a mechanism to support incremental innovation in developing economies are needed to reduce costs and increase technological diffusion.

2.1.4 Alternative Measurements

Innovation extends beyond patents and requires complementary indicators such as R&D expenditure, publications, and adoption metrics. These alternative measures help capture different stages of the innovation lifecycle and are crucial for assessing non-technological and incremental innovations, especially in developing contexts.

In order to provide a more comprehensive perspective on the measurement of technologies, this section will present a simplified overview of alternative methods for identifying and measuring technological progress. Haščič and Migotto (2015) summarize the most common methods for measuring technology development (see Table 3). Beyond patents, the table offers a structured overview of alternative measures of novel technologies, categorized by different stages: technology development, diffusion, adoption, and non-technological innovations. Each of these innovation measures possesses a unique set of strengths and limitations, which affect their suitability for various research and policy applications. While certain indicators, such as patent data, offer extensive historical records and classification possibilities, they may not encompass all aspects of innovation, particularly non-technological advancements. Conversely, other measures, including licensing surveys and

market penetration data, provide direct insights into technology adoption but are constrained by issues related to data availability and confidentiality.

Table 3*: Alternative measures of innovation and their key features.*

Stage of Innovation	Measures	Advantages	Limitations
Technology Development	**R&D expenditures and personnel**	- Easy to communicate and compare across sectors - Provides insight into investment in innovation	- Measures inputs rather than actual innovation outcomes - Difficult to distinguish environmental activities - Limited data availability, mostly restricted to OECD countries and certain industries
	Scientific publications	- Broad geographical and historical coverage - Can capture some aspects of environmental innovation	- Not all research leads to technological applications - Identifying environmental relevance can be challenging
	Patented inventions	- Direct measure of technological innovation - Captures intermediate innovation outputs - Allows detailed classification, including environmental aspects - Global database with long historical records	- Excludes non-technological innovations - Delays between invention and patent approval can affect timeliness
Technology Diffusion	**Patenting activity & International trade**	- Can track the spread of new technologies across borders	- Difficult to identify environmentally relevant products - Many traded goods are not necessarily innovative
Technology Adoption	**Licensing surveys**	- Reflects the commercial value of innovations through royalties	- High costs and confidentiality concerns limit data availability
	Sales and market penetration	- Indicates real-world impact on environmental outcomes	- Data can be difficult to obtain due to confidentiality and reporting inconsistencies
Non-Technological Innovations	**Innovation surveys**	- Can capture changes in business processes and management practices - Useful for understanding broader innovation trends	- Expensive and resource-intensive to conduct - Data comparability across regions and industries can be problematic

Source: Haščič, I., & Migotto, M. (2015). Measuring environmental innovation using patent data. OECD Environment Working Papers No. 89.

2.2 Green Patents

While patents in general serve as indicators of technological advancement, green patents specifically focus on innovations that contribute to environmental sustainability. Kraus et al. (2020) define green innovations as technological developments designated to reduce waste, mitigate global warming, decrease water and air pollution, minimize reliance on coal, oil, and electricity, and enhance energy conservation. Given the urgency of climate change mitigation, understanding the trends, impact, and diffusion of green patents is essential. The next section explores how green patents are classified and measured in existing research.

2.2.1 Green Classification

Multiple classification systems—ENV-TECH, IPC Green Inventory, and Y02—are used to identify green patents, each with varying levels of accuracy and applicability. Y02 stands out for its balance between usability and specificity, making it particularly useful for both policy and research applications.

According to Favot et al. (2023), four methods are currently employed to identify green innovations in patents: classification by codes, classification by keywords, a combination of the previous two methods, and a manual selection process. The most common method is to use codes presented in databases published by several international organizations mentioned earlier.

In their study, Favot et al. (2023) examine three methods for identifying green innovations using IPC/CPC classifications:

- The first method was developed by the OECD and is called *environment-related technologies* (ENV-TECH). ENV-TECH's advantages are that it uses both IPC/CPC classifications and that it covers a wide variety of green innovations in multiple industries.
- The second method is the *IPC Green Inventory*, developed by WIPO with the aim of identifying environmental sound technologies (ESTs) through IPC codes.
- The third method is the *Y02 Tagging Scheme*, developed in a collaboration between the EPO and the USPTO to identify climate change mitigation technologies (CCMTs) through CPC codes.

A comparative evaluation of these methodologies reveals significant differences in their effectiveness and coverage (Favot et al., 2023); a summary of the comparison between the different methods is presented in Table 4. ENV-TECH provides the broadest classification, capturing 82.41% of green patents when tested on a dataset of EPO applications. In contrast, the IPC Green Inventory identifies 42.76% of green patents, while the Y02 Tagging Scheme captures 67.97% (Favot et al., 2023). The overlap among the three methods was limited, with only 22.47% of patents being identified by all three methodologies, indicating that each classification system captures unique aspects of green innovation (Favot et al., 2023).

Table 4: *Comparison of the Three Methods.*

Method	Organization	Codes Used	Focus Area	Strengths	Limitations	Accuracy
ENV-TECH	OECD	IPC + CPC	Broad environmental tech (e.g., energy, waste, transportation)	Covers the widest range of patents, frequently updated	Can be redundant, some duplication	82.41%
IPC Green Inventory	WIPO	IPC	Environmental sound technologies (ESTs)	Well-established, used in global patent studies	Limited to IPC codes, less adaptable	47.46%
Y02 Tagging Scheme	EPO	CPC	Climate change mitigation (CCMTs)	Regularly updated, specific to low-carbon innovation	Limited to CPC, identifies fewer patents	67.97%

Source: Favot, M., Vesnic, L., Priore, R., Bincoletto, A., & Morea, F. (2023). *Green patents and green codes: How different methodologies lead to different results*. Resources, Conservation & Recycling Advances, 18, 200132.

The Y02 tagging scheme shows a lower accuracy (about 14% lower) than ENV-TECH in the study but is significantly better (about 20% higher) than the IPC Green Inventory. Rainville et al. (2025) have in a similar manner examined the ability of patent classifications to effectively capture innovation in green technologies. Their findings indicate that the Y02 method outperforms the IPC Green Inventory in terms of capturing green innovation and making fewer misclassifications.

Haščič and Migotto (2015) highlight other advantages of the Y02 method that are not considered by Favot et al. (2023). The authors argue that the Y02 tagging scheme has significant relevance not only for the business community but also for the research and policy community. The process of identifying green patents can be complex and time-consuming; the Y02 tagging scheme helps even inexperienced actors to identify climate change mitigation

technologies. Haščič and Migotto (2015) posit that the identification of green innovation through the ENV-TECH methodology necessitates a search strategy, which involves a comprehensive analysis of the extant literature in each technological domain. This analysis entails the identification of a suitable IPC/CPC code that meets the criteria in the field, as well as the exclusion of irrelevant patents from the selection process. The authors further posit that two common errors occur during the initiation of the search strategy: (i) the inclusion of irrelevant patents and (ii) the exclusion of relevant ones. Haščič and Migotto (2015) contend that the integration of the Y02 scheme within the CPC classification serves to streamline the search process, where non-experts with limited resources could utilize the method for green patent identification.

2.2.2 Trends in Green Patent Research

Empirical studies show that green patents are positively linked to firm performance, national competitiveness, and sustainability transitions. However, the relationship between green innovation and regulatory or economic conditions varies across regions and income levels, indicating the importance of context-sensitive policy design.

Research on green patents has been conducted in various industries, and it has become an important part of balancing activities between economic factors, political decisions, regulations, or laws, and subsequently the long-term competitiveness of companies or countries. As delineated in Table 4, which details alternative methods for measuring innovation, the development phase of novel technologies is typically assessed through the following metrics: research and development (R&D) expenditures, scientific publications, and registered patents (Haščič & Migotto, 2015). Green innovation aligns with this prevailing paradigm and is often quantified by the allocation of R&D expenditures to green innovation or the number of registered green patents.

Urbaniec et al. (2021) have examined trends in patent applications, focusing on green technological innovations between the years 2000 and 2017. Their analysis revealed an increase in patent applications in areas such as climate change mitigation technologies (CCMT) and green energy. The authors argue that this trend can be attributed to the impact of environmental policies on actors' investments and determination to meet sustainability goals.

Scarpenellini et al. (2019) conducted a study to examine the financial performance of firms that prioritize environmental sustainability or possess registered green patents. The

study's findings indicate that engagement in green innovation is associated with enhanced financial performance. This suggests that green patents, in conjunction with R&D intensity, can serve as an effective proxy of firm success. In a similar vein, Martin-Vinuesa et al. (2020) have employed the resource-based view (RBV) theory and the partial least squares (PLS) method to examine the relationship between green investments and firm success. Their findings indicate that green investments contribute to stability and long-term competitiveness.

A substantial body of research has been dedicated to examining the motivation behind economic incentives and political policies regarding green innovation (Chang & Cheng, 2024; Hall, 2009; Serener et al., 2023). De Lange (2024) examined the impact of green patents on countries' GDP over time and the potential for government policies to accelerate the transition towards renewable energy. The study revealed a positive correlation between green innovation and countries' GDP over a decade. Moreover, Cai and Xu (2021) assess the impact of China's New Environmental Protection Law (NEPL) on eco-innovation among publicly traded corporations. Utilizing a difference-in-difference (DID) methodology, the study finds that the NEPL has a substantial restraining effect on enterprise green innovation, particularly in the domain of high-degree invention patents. The primary reason for this decline is the tightening of financial constraints imposed by stringent environmental regulations. The study further suggests that although regulatory pressure can promote environmental governance, it may also impede innovation when financing becomes challenging (Cai & Xu, 2021).

A divergent approach is adopted by Du, Li, and Yan (2019), who undertake an examination of the relationship between green technology innovations and carbon emissions utilizing patent data from 71 economies from 1996 to 2012. Their empirical analysis identifies a threshold effect; whereby green technology innovations contribute to emission reductions exclusively when economies attain a high-income level.

Contrary to the prevailing perspective of examining green patents as an indicator of reduced carbon emissions, Wang et al. (2020) have investigated the possibility that carbon emissions are the driving force behind green investments and innovations. The authors of this study posit that carbon emissions act as a catalyst for environmental legislation, which, in turn, fosters positive incentives for green innovations. Specifically, they contend that carbon dioxide exerts a direct influence on transportation technology patents classified under Y02T, while concurrently acting as an indirect catalyst for technologies dedicated to carbon capture and storage (Y02C). The authors underscore the dual impact of carbon emissions, highlighting their detrimental effect on the natural environment and their role as a catalyst for green innovation. They emphasize the necessity for substantial government policy support to nurture this process.

3. Theoretical Framework

This chapter presents the theoretical underpinnings that guide the analytical design of the study. The framework integrates innovation diffusion theory, particularly the logistic S-curve model, with established research on technology life cycles and patent analysis. The S-curve is a methodological framework employed to analyze the trajectory of technological maturity based on longitudinal patenting activity. This framework provides a structured lens through which trends in green innovation can be assessed.

3.1 Technological Uncertainty

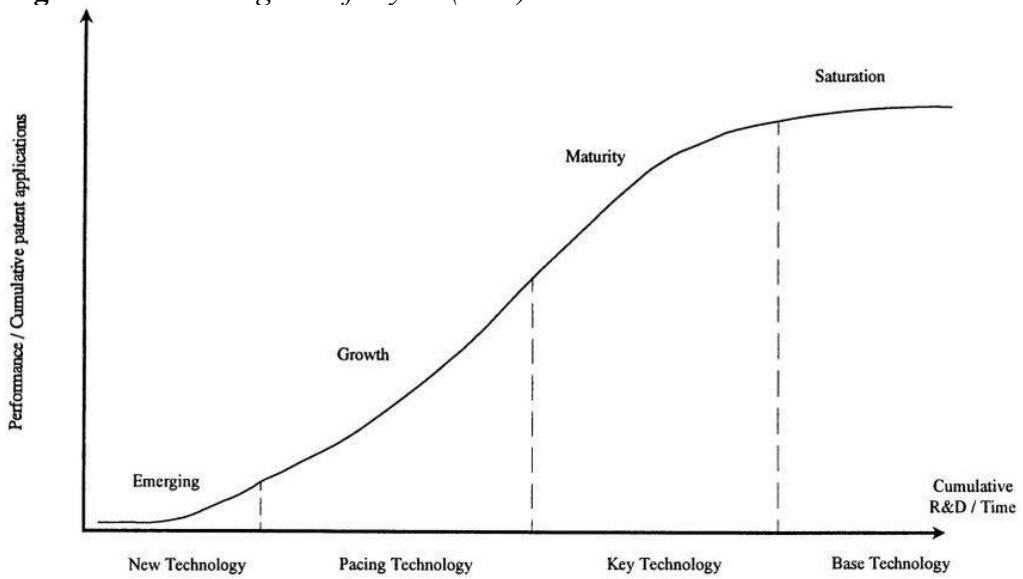
Innovation and technological development are inherently unpredictable and characterized by high uncertainty, making it challenging to predict whether innovation ambitions will succeed or fail. In addition to the evident uncertainty inherent in innovation, which pertains to the development of the product itself, Fleming (2001) underscores the presence of uncertainties concerning the adoption and diffusion of the novel technology, the preferences and acceptance of markets and customers, and the actions and strategies employed by competitors. The concept of innovation can be fundamentally explained through a recombinant process, whereby novel technologies are developed by combining existing components in novel ways (Fleming, 2001). The author contends that the uncertainty of the innovation is contingent on the strategy of recombination selected by the inventor. The initial strategy involves the combination of familiar components, which reduces variety and novelty, resulting in incremental innovation that is more predictable. The second strategy involves experimenting with novel or unfamiliar components, which leads to increased variability and an increase in uncertainty with more failures, but successful endeavor leads to breakthrough innovations (Fleming, 2001).

3.2 The S-Curve and Patent-Based Maturity Assessment

Another approach is to examine how innovation and its uncertainty evolve over time. Klepper (1996) describes the development of the product life cycle (PLC), where novel products are initially offered in a variety by many players as entry markets. Initially, the level of product innovation is high, but as the market expands, a critical juncture is attained where market exit becomes more prevalent than market entry (Utterback and Abernathy, 1975).

Consequently, the emphasis shifts from ongoing innovation to streamlining processes, reducing the variety of products offered by a select group of players. Mueller and Tilton's (1969) case study concluded that the introduction of a new product implied high uncertainty about customer preferences. Consequently, the product is presented in different variations, and companies focus on research and development. Furthermore, the authors posit that over time, a predominant design will emerge, resulting in a shift in focus from innovation to efficient production. This technological discontinuity follows an S-shaped curve (Mueller & Tilton, 1969; Tushman & Anderson, 1986).

Figure 3*: Technological life cycle (TLC) and the S-curve.*



Source: Gao et al. (2013). Technology life cycle analysis method based on patent documents. Technological Forecasting and Social Change. 80. 398–407. 10.1016/j.techfore.2012.10.003.

Investing in technology and the associated patent application process is a costly and time-consuming endeavor. The investment potential of a technology is contingent, to a considerable extent, on the stage of its life cycle (Haupt et al., 2007). A prevalent method for examining technology life cycles involves the analysis of patent applications (Haupt et al., 2007). Empirical studies demonstrate an S-shaped evolution of the number of patent applications (Haupt et al., 2007). In a manner consistent with the understanding of product life cycles, the technology life cycle (TLC) can be conceptualized as comprising an introduction stage, a growth stage, a maturity stage, and a decline stage (Haupt et al., 2007). The utilization of patent documentation as a foundation for TLC representations possesses numerous benefits when compared to conventional metrics, such as accumulated product sales (Haupt et al., 2007). First, patents provide information about the technological development itself, as they describe the technical part of the invention. Secondly, patent documentation contributes to

information about the commercial potential of a technology, as it is one of the preconditions for patentability. Finally, the evaluation of patent applications can be conducted on an annual basis over time using databases (Haupt et al., 2007). This approach facilitates a systematic and objective assessment of the progress and trends in patent applications. This progression unfolds through distinct phases: introduction, growth, maturity, and, ultimately, saturation or decline (Sossa et al., 2016).

In light of the TLC, determining where a particular technology is on its S-curve is an important challenge. In the domain of innovation research, patent data have emerged as a pivotal proxy for technological activity and progress. Patents, by their very nature, serve as a tangible testament to the fruits of inventive labor, with the accumulation of these intellectual properties over time providing a quantifiable metric for assessing the rate of technological innovation within a given society or economy. Researchers have observed that the cumulative number of patent applications for a given technology typically follows an S-shaped curve (Haupt et al., 2007). However, the method is not without significant limitations. Haupt et al. (2007) emphasize that users must consider the necessity of a comprehensive list of patents within a particular technological domain. This assertion is substantiated by the documented challenges associated with patent shortcomings (Allison et al., 2003; Sichelman, 2009; Mansfield, 1986; Hall et al., 2000) and the inherent difficulty of patent offices in achieving comprehensive classification (Li et al., 2018).

Gao et al. (2013) emphasize that the assessment of a technology's life cycle stage entails the observation of its patent application trends over time and the fitting of these trends to an S-curve. An escalating number of patents is indicative of an escalating degree of inventiveness, which is a feature of the growth stage, while a plateau in patenting indicates that the technology is reaching maturity or saturation. This approach utilizes patent metrics as a quantitative representation of innovation performance. However, Gao et al. (2013) also caution against an overreliance on the number of patents, suggesting that such an approach may offer an overly simplistic understanding of the complex reality. Patent trends may lag actual innovation or be influenced by policy and strategic behavior; therefore, a pure count-based S-curve should be interpreted cautiously.

4. Method

The method chapter begins by presenting the research strategy and design, detailing the epistemological stance, the use of secondary patent data, and the rationale for selecting specific CPC classifications. The following section provides a detailed overview of the tools and libraries employed for data management, modeling, and visualization. The following sections proceed to methodically examine each phase of the analysis, starting from the initial data preprocessing and logistic curve modeling and culminating in the implementation and validation of Latent Dirichlet Allocation (LDA) for topic modeling. The chapter introduces complementary visualization methods (LDAvis and t-SNE) that serve to enhance interpretability and reveal document-level relationships. The chapter's concluding remarks offer a critical evaluation of research quality criteria, encompassing reliability, validity, objectivity, and replicability. This evaluation is accompanied by an acknowledgment of the study's methodological limitations.

4.1 Research Strategy

This study employs a quantitative and exploratory research approach, leveraging large-scale textual data to identify thematic patterns in green patents. Quantitative studies frequently employ hypothesis testing and causal inference methods; however, this research aligns more closely with Bryman and Bell's (2019) characterization of a descriptive, data-driven approach to quantitative analysis. Rather than testing specific hypotheses, the objective is to identify latent semantic structures, manifesting as topics and clusters, within the text of climate-related patent titles and abstracts. These structures are interpreted as reflections of distinct technical domains. The research employs an inductive approach to reasoning, deriving patterns from data without a predetermined theoretical model. Nevertheless, the theoretical framework of the s-curve is utilized to identify and justify the selection of a technological domain (i.e., CPC classification) for further thematic study. The selection of datasets is informed by existing classifications of climate change mitigation technologies (CPC Y02/Y04).

While the study's methodology is not qualitative, it shares with qualitative research the intention to generate insights from observed patterns rather than to confirm or falsify hypotheses (Bryman & Bell, 2019). The study's quantitative orientation is attributable to its reliance on computational models, statistical techniques, and large-scale document processing.

The study's epistemology aligns with positivism, emphasizing observable, objective patterns in text, and its ontology follows an objectivist perspective, positing the existence of thematic structures in data that can be discerned through algorithmic techniques (Bryman & Bell, 2019).

4.2 Research Design

The study is designed as an exploratory cross-sectional analysis using a single-point-in-time export of 50,000 patents. Although the patent documents themselves span a broad historical range (1896-2025), the analytical treatment is static and non-comparative, consistent with the logic of cross-sectional design (Bryman & Bell, 2019).

The research design is primarily descriptive, addressing questions such as, what themes or topics dominate the landscape of green innovation patents? How can these topics be grouped into meaningful clusters? What trends or developments can be inferred from the occurrence of these themes over time? There is no attempt to investigate causal links between topics and performance indicators, such as citations or commercial success. Instead, the focus is on mapping the thematic terrain of green patents and creating a structured representation of the innovation space using unsupervised machine learning techniques.

This is consistent with the use of secondary data for discovery-oriented purposes (Bryman & Bell, 2019), where researchers use already existing data sources to reveal new insights, often in an inductive and flexible way. By combining topic modeling, clustering, and visualization, the design facilitates a multidimensional interpretation of technical domains and semantic shifts in climate-related patenting.

4.2.1 Literature Collection

In addition to the primary analysis of patent data, this study incorporates a selective review of secondary academic and institutional sources to provide information on the interpretation of the topic and the contextual framing. The literature review was conducted using *Scopus* as the primary academic database due to its extensive indexing of peer-reviewed research in innovation studies, technology policy, and environmental management. To identify relevant literature not covered by *Scopus*, additional searches were conducted on *Google Scholar,* with a focus on topic modeling applications and patent analysis.

To strengthen the environmental and policy relevance of the analysis, documents from international organizations such as the Intergovernmental Panel on Climate Change (IPCC) and the United Nations Framework Convention on Climate Change (UNFCCC) were

consulted. These reports were instrumental in contextualizing the green patent domains within a broader societal and technical framework, thereby ensuring that the extracted topics were interpreted with reference to recognized priorities for reducing climate impacts.

The literature review was guided by targeted keyword searches designed to capture academic and technical work relevant to green patents, topic modeling, and data-driven innovation analysis. The search terms included combinations of: *Green innovation, green technology, sustainable innovation; Patent analysis, patent citations, patent data mining; Latent Dirichlet Allocation, topic modeling, unsupervised learning, text mining*, where Boolean operators (e.g., AND/OR) were used. The application of these search terms resulted in the identification of a substantial number of articles, which were subsequently exported from the Scopus database. The identified literature was subsequently managed in an Excel spreadsheet, and the abstracts were reviewed to identify literature relevant to the study. The application of a green mark to articles deemed relevant facilitated the identification of articles requiring more thorough review.

4.2.2 Tools and Software

All data preprocessing, topic modeling, clustering, and visualization were carried out using Python version 3.12, executed through the Spyder integrated development environment (IDE) within the Anaconda Navigator distribution. This environment was selected for its accessibility, built-in support for data science packages, and suitability for researchers with limited prior coding experience.

Table 5: *Summary of the Tools Used from Python Libraries.*

Library/Tool	Purpose
pandas, NumPy	Data handling and structuring
scikit-learn	Preprocessing, clustering (k-means), evaluation metrics (e.g., silhouette score)
NLTK	Natural language preprocessing and stopword filtering
Gensim	Latent Dirichlet Allocation (LDA) topic modeling
pyLDAvis	Interactive topic visualization
matplotlib, seaborn	General data visualization
t-SNE	Dimensionality reduction and topic projection

Note: *Table 5 outlines the Python-based tools and libraries employed throughout the data analysis pipeline. These tools enabled efficient preprocessing, topic modeling, clustering, and visualization. Each library contributed a distinct functionality—from data manipulation using pandas and NumPy to topic modeling via Gensim.*

Coding support and troubleshooting were guided by official documentation (e.g., scikit-learn) and example repositories on GitHub. This reliance on community and official resources ensured transparency and reproducibility while also supporting the author's learning curve in implementing computational models.

Additionally, *ChatGPT* from OpenAI was also used as coding support and a troubleshooting guide. Due to my limited coding experience, this generative AI was used as a convenient way to debug code and explain what mistakes were made in certain executions.

4.3 Data Collection

For this study, Lens.org has been selected as the patent database due to its comprehensive coverage and accessibility. Lens is an open-access database that provides structured patent data from multiple jurisdictions, including the United States Patent and Trademark Office (USPTO), the European Patent Office (EPO), and the World Intellectual Property Organization (WIPO) (Lens, 2025a). The selection of Lens.org is further informed by its advanced search functionalities, encompassing Boolean logic, structured search, biological search, classification-based search, and sorting filters, thereby facilitating precise and efficient patent retrieval (Lens, 2025b). In contrast to proprietary databases that impose financial and access constraints, Lens.org guarantees that all users have unrestricted access to crucial patent knowledge.

The initial phase of data collection entailed the establishment of a free account on Lens.org, followed by the utilization of the database search function to identify green patents. The selection of the Y02 tagging scheme is motivated by two key objectives: first, to optimize efficiency by employing a less sophisticated search strategy, and second, to minimize the occurrence of both type I and type II errors. This is achieved by strategically incorporating or omitting relevant patents in the survey, as previously outlined in the work of Haščič and Migotto (2015). Given the objective of this study, which was to examine trends in green technology, CPC codes were deemed more suitable than creating a query with keywords related to green technology. The patent collection subsequently concentrates on identifying all patents that contain any of the CPC codes created by the USPTO and EPO (refer to Table 6 for an overview of CPC section Y and a description of which technologies it covers).

Table 6: *Overview of CPC Section Y.*

Main Class	Description
Y02	**Technologies or applications for mitigation or adaptation against climate change**
Y02A	Technologies for adaptation to climate change (human, industrial, and economic activities)
Y02B	Climate change mitigation technologies related to buildings (housing, appliances, end-user applications)
Y02C	Capture, storage, sequestration, or disposal of greenhouse gases (GHG)
Y02D	Climate change mitigation technologies in information and communication technologies (ICT) aimed at reducing energy use
Y02E	Reduction of GHG emissions related to energy generation, transmission, or distribution
Y02P	Climate change mitigation technologies in the production or processing of goods, including agriculture and fishing
Y02T	Climate change mitigation technologies related to transportation
Y02W	Climate change mitigation technologies related to wastewater treatment or waste management
Y04	**Information or communication technologies impacting other technology areas**
Y04S	Systems integrating power network operation, communication, or information technologies for smart grids
Y10	**Technical subjects covered by former USPC cross-reference art collections (XRACs) and digests**
Y10S	Technical subjects formerly covered by USPC cross-reference art collections (XRACs) and digests
Y10T	Technical subjects formerly covered by US classification

Source: United States Patent and Trademark Office (USPTO). (n.d.). *COOPERATIVE PATENT CLASSIFICATION*. Classification Resources.

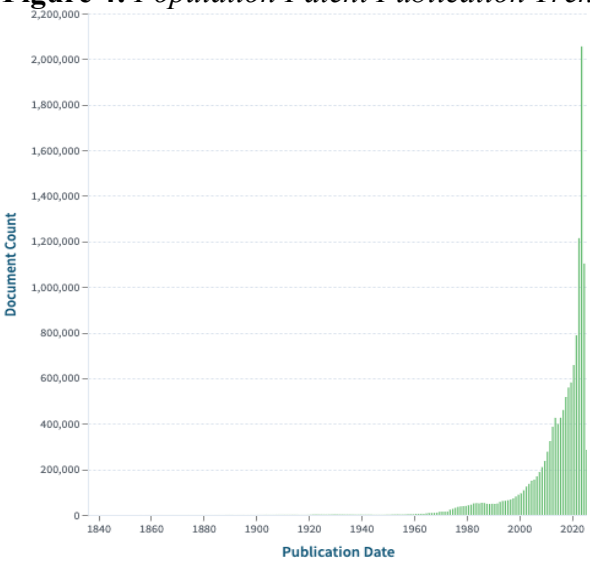
To identify green patents, the Lens database search function was utilized, resulting in the formulation of the following query: *"CPC classification code: Y02* OR CPC classification code: Y04*."* This search query yielded 13,591,586 patents, which were subsequently stored in a collection within the Lens account. The patents could then be exported according to the following criteria: the *number of documents* to include, the f*ile format*, and the choice of *patent information.*

In the context of patent exportation, the maximum number of patents that could be selected for download was set at 50,000. Despite exerting considerable effort, I have been unable to ascertain the distribution or the selection of the maximum number (n = 50,000) of patents upon downloading. Consequently, it is assumed that the distribution is random. The

assumption of randomness will be supported by a comparison of distribution between the total population and the sample dataset.

Figure 4 presents a graphical representation of the distribution over time for the total population of green patents via Lens (n=13,591,586). The earliest document references 1836 (n=1), while the most recent pertains to 2025 (n=284,120). The trend demonstrates an increase from the beginning of the 2000s and an accelerated growth rate in the 2010s, with the highest recorded value in 2023 (n= 2,055,383). The analysis via Lens' website further suggests that China has applied for the most patents (6,373,778 or 47%), followed by Japan (n=1,594,331 or 12%) and the US (n=1,378,607 or 10%). The most prevalent IPC codes in the search query were H01MB/04 (n = 214,159), followed by H01M10/0525 (n = 206,409). The most prevalent CPC codes in the search query were Y02E60/10 (n=1,657,869), followed by Y02P70/50 (n=701,813). This descriptive data will then be compared with the data set to support the argument that the distribution of patents follows a reasonably random and representative sample with regard to the maximum number of 50,000 patents exported from the database (see section 3.1.1 Data Overview).

Figure 4: *Population Patent Publication Trend from 1836 to 2025 (n=13,591,586).*



Source: *https://www.lens.org/lens/search/patent/analysis?collectionId=231260*

Note: *This histogram illustrates the annual distribution of patent documents (n = 13,591,586) retrieved from the Lens.org database, covering publication years from 1836 to 2025. The data shows a gradual increase in patenting activity throughout the 20th century, followed by an exponential surge in the 21st century, with the highest volume recorded in 2023.*

The next step in the data retrieval process was the selection of a file format. The CVS file format was selected for this export, and the following patent information was systematically reviewed for inclusion: *#; Jurisdiction; Lens ID; Year of Publication; Title;*

Abstract; Applicant; Inventor; Cited Patent Count; Cited by patent count; Simple family size; Extended family size; CPC classifications; IPCR classifications; Legal status. An exemplar of one of the patents exported, along with pertinent information, is presented in Table 7.

Table 7*: Example of a Patent Included in the Dataset.*

#	Jurisdiction	Lens ID	Publication Year	Title	Abstract	Applicants	Inventors
31	US	198-708-985-540-33X	2013	Energy systems, energy devices, energy utilization methods, and energy transfer methods	****	HAMILTON SCOTT ROBERT;;DEMAND ENERGY NETWORKS INC	HAMILTON SCOTT ROBERT

Cites Patent Count	Cited by Patent Count	Simple Family Size	Extended Family Size	CPC Classifications	IPCR Classifications	Legal Status
108	4	14	14	H02P9/04;;H02J3/32;;H02J3/381;;H02J3/46;;H02J2300/24;;H02J2300/28;;H02J2300/40;;H02P9/46;;Y02E10/56;;Y02E10/76;;H02J3/28;;H02J11/00;;Y02E70/30	H02P9/04;;H02J7/00	ACTIVE

Note: *Table 7 provides a detailed overview of a representative patent included in the dataset. It highlights key metadata such as jurisdiction, applicant and inventor identities, citation metrics, family size (indicating geographical protection breadth), and classification under both CPC and IPCR systems.* **** *= Abstract has been purposely removed in this example due to its extensive length.*

In this particular example, it is evident that the patent in question was issued in the United States in 2013 and is still active. According to the title, the patent pertains to energy solutions for use and transfer. The patent was applied for by an inventor named Robert Scott Hamilton, along with the company Demand Energy Networks INC. The patent makes reference to 108 previous patents and has been cited by 4 more recent applications. The patent is comprised of 14 family sizes, which indicates broad jurisdiction protection. The patent in question has been assigned several different CPC codes, three of which fall within the classification Y02E, which is concerned with the reduction of greenhouse gas (GHG) emissions in relation to energy production, transmission, or distribution. This constitutes a broad CPC subclass under the broader category Y02, which encompasses inventions directed towards climate change mitigation technologies, with a particular focus on those relevant to the energy sector. The subsequent Table 8 presents the three subclasses identified within the patent, along with their respective descriptions.

Table 8: *Green CPC Codes Identified in Example Patent from Database.*

CPC Code	Description
Y02E10/56	Power conversion systems, e.g. maximum power point trackers
Y02E10/76	Systems integrating renewable energy sources and electric power distribution
Y02E70/30	Systems combining energy storage with energy generation of non-fossil origin

Note: *Table 8 lists the green CPC codes assigned to a selected example patent focused on energy management and storage solutions. Each CPC code corresponds to a specific subdomain within climate change mitigation technologies. These classifications reveal the patent's emphasis on renewable energy integration, power conversion, and non-fossil energy storage.*

3.1.1 Data Overview

To get a better understanding of the downloaded dataset, the following section will provide an overview of the descriptive statistics. Table 9 presents the descriptive statistics of the numerical variables found in the dataset (n = 50,000). Furthermore, the data represents patents from 1896 to 2025 (mean = 2012.72, sd = 11.30). An examination of the number of citations reveals that both backward and forward citations have similar means (4.71 and 4.17, respectively). It is also evident that most patents have few or near-zero citations, suggesting that few patents have a high impact, which aligns with claims earlier presented by Arts et al. (2013). The mean family size for simple families is 6.64, while extended families exhibit a notably higher average of 9.39. The median values also exhibited a similar trend, with both measuring 3.00, indicating that most patents belong to relatively small families. However, the standard deviation was found to be significantly higher for extended families (29.75) than for simple families (13.05), suggesting a wider spread and the presence of very large patent families in the extended group.

Table 9: *Descriptive Statistics of Numerical Variables.*

Statistic	Publication Year	Backward Citations	Forward Citations	Simple Family Size	Extended Family Size
Count	49,999	50,000	50,000	50,000	50,000
Mean	2012.72	4.71	4.17	6.64	9.39
Std	11.30	23.21	17.09	13.05	29.75
Min	1896	0.00	0.00	1.00	1.00
25%	2010	0.00	0.00	1.00	1.00

50%	2016	0.00	0.00	3.00	3.00
75%	2020	5.00	2.00	8.00	9.00
Max	2025	2,277.00	786.00	393.00	730.00

Note: *Table 9 presents descriptive statistics for key numerical variables within the patent dataset. It includes information on publication years, citation metrics, and patent family sizes. Citation counts—both backward and forward—exhibit significant skewness, indicating the presence of a few highly cited patents. Family size statistics suggest that while many patents belong to small families, some have extensive international filings. These metrics offer insight into the temporal distribution, influence, and legal scope of the patents analyzed.*

In a similar manner, descriptive statistics are presented in Table 10, which contains non-numeric variables from the dataset. These statistics include unique values, peak values, and their frequency, as well as missing values within the dataset. The dataset under consideration contains 72 unique jurisdictions (i.e., the country or region where the patent is legally processed), with China (CN) being the most frequent, representing 18,746 patents. The applicant field contains 23,746 unique entities, and in terms of inventors, 28,328 individuals are represented. The CPC field with Y02E60/10—"Energy storage using batteries"—being the most frequent classification (n=7,211). The IPCR with H02J3/38 being the most prevalent classification (n=1,599). The legal status column comprises seven distinct statuses, with most patents designated as active (n=16,641), signifying that a substantial proportion of patents are currently in effect.

Table 10: *Descriptive Statistics of Non-Numerical Variables.*

Column	Unique Values	Top Value	Frequency
Jurisdiction	72	CN (China)	18,746
Applicants	23,746	GS YUASA INT LTD	560
Inventors	28,328	WOBBEN ALOYS	154
CPC Classification	16,428	Y02E60/10	7,211
IPCR Classification	27,821	H02J3/38	1,599
Legal Status	7	ACTIVE	16,641

Column	Non-null Count	Missing values (%)
Abstract	40,961	~18% missing
Applicants	49,708	~0.6% missing
Inventors	48,151	~4% missing
CPC Classification	50,000	0% missing
IPCR Classifications	49,158	~1.7% missing

Note: *Table 10 provides a descriptive overview of the non-numerical variables in the patent dataset. The top section summarizes the number of unique entries per column, along with the most frequent value and its count. The bottom section details the completeness of the data, showing the number of non-null entries and the percentage of missing values for key fields such as abstracts, applicants, and inventors. This summary informs the overall reliability and representativeness of the dataset used for topic modeling.*

To validate the representativeness of the sample of 50,000 patents used in this study, descriptive statistics were compared with the broader dataset obtained from the original search via Lens.org (n = 13,591,586). The complete dataset encompasses patent registrations from 1836 to 2025, with a notable surge after 2010 and a peak in 2023 (n = 2,055,383). In a similar vein, the selected dataset extends from 1896 to 2025, with an average publication year of 2012.72 (SD = 11.30), thereby mirroring the observed exponential growth across the entire corpus. The distribution of jurisdiction is consistent across both datasets. In the complete dataset, China accounts for 47% of patents, and in the sample dataset, China (CN) is again the most common jurisdiction (n=18,746 or 37.49%), suggesting that the sample is reasonably consistent with global patterns of patent activity. Moreover, the most prevalent CPC code in both datasets is Y02E60/10, which denotes energy storage utilizing batteries. In the complete dataset, Y02E60/10 accounted for approximately 12.2% of patents (n=1,657,869) and in the sample dataset, 14.4% (n=7,211). Non-numerical variables also demonstrate diversity and relevance. The sample encompasses 72 jurisdictions, over 23,000 unique applicants, and more than 28,000 inventors. The comparison yielded a confirmation that the selected dataset can be regarded as acceptable and reasonably consistent with the broader trends in the entire dataset with respect to temporal, jurisdictional, and technological distribution. It can be posited that the sample under consideration provides a sufficiently representative cross-section of the global green patent landscape.

4.4 Data Analysis

This chapter outlines the full analytical process undertaken to uncover latent themes within green technology patents, with a particular focus on the CPC subclass Y02E60/10. The analysis begins with a comprehensive data cleaning and preprocessing routine to ensure textual consistency and semantic clarity. To identify promising technological areas, logistic S-curve modeling is applied to assess growth patterns in patent activity. The core analytical method, Latent Dirichlet Allocation (LDA), is then used to extract hidden thematic structures from patent abstracts and titles. Model parameters are carefully tuned and validated using coherence scores. The interpretability of the LDA results is further enhanced through visualization tools such as LDAvis and t-SNE, which support both semantic inspection and structural clustering.

4.4.1 Data Cleaning and Preprocessing

To ensure data quality and relevance, several filtering and cleaning methods were applied. Text preprocessing is an important preparatory step that lays the foundation for meaningful topic modeling and ensures that the model can identify patterns and thematic structures more effectively (Pan and Xue, 20).

The initial phase of the data cleansing procedure entailed the selection of a more reasonable time span. This was necessitated by the findings of the data overview, which identified the earliest patent within the CPC class Y02 as early as 1896. This period is not known for prioritizing or putting a large societal emphasis on eco-friendly inventions. To capture the advent of environmentally sustainable inventions, a more limited time span was selected, ranging from 1980 to 2024, thereby excluding the 127 patents filed in 2025. Moreover, the dataset was reduced to 48,803 patents, indicating that 1,070 patents predating 1980 were excluded.

Subsequently, a series of text preprocessing actions were implemented to ensure the quality of the LDA. Initially, the titles and abstract fields were combined into a single text field, thereby incorporating both into the construction of the LDA model. Thereafter, the text was cleaned of numerals and punctuation, and all letters were converted to lowercase. Given the international nature of the dataset, which encompasses patent documentation from various countries and classification organizations, the text contained languages other than English. To address this, non-ASCII characters were eliminated, which is equivalent to non-Latin

alphabets. This category includes Chinese characters, accented letters, and symbols. The impact of non-English patents will be presented in the result section as "language noise impact," which is defined as the number of patents affected from the sample of the technical domain chosen for LDA modeling.

The following step entailed the removal of stop words, i.e., words such as "the," "are," or "but." This step involved the utilization of the sklearn library (ENGLISH_STOP_WORDS), which contains 318 prevalent English stop words. Additionally, word clouds were employed to identify "domain stop words," defined as words frequently used in patents but not specific to any technology or field, such as "method," "device," and "unit." The analysis of word clouds led to the identification of 66 domain stop words. The aim of stop word removal was to improve interpretability by removing terms that are functionally repetitive but conceptually uninformative.

To improve interpretation and capture frequently occurring multi-word expressions, bigrams were constructed. The use of bigrams implies two adjacent words that occur more frequently than would be expected by chance (Source). Using the Gensim library's Phraser models, bigrams were automatically detected based on their frequency and statistical association across the corpus. A minimum number (10) and a threshold (10) were defined to ensure that only meaningful and commonly occurring word pairs were retained. These bigrams were then treated as single tokens (e.g., “energy storage” becomes “energy_storage”), which allows the topic model to recognize compound technical concepts that are essential in patent language.

Finally, the nltk's ‘SnowballStemmer’ was employed to transform words into their stems, e.g., "change," "changing," and "changed" stem to "chang." Stemming can be valuable for grouping variations of an expression, which can increase the reliability of clustering by reducing vocabulary and potentially illustrating clearer patterns (Source). However, the downside can be the difficulty of interpreting the stem word and its semantic meaning (Source). To address this challenge, a comparison will be made between the stemming method and a stemming-free approach. This comparison will ensure that the thematic relevance of keywords and their interpretation are adequately addressed.

4.4.2 Identify Growth CPC-codes using Logistic S-Curve

The objective of this study was to examine patents that are of interest, given the continuous growth in the number of applications. The technological domain has not yet reached a state of stagnation, characterized by a dominant design that has stabilized and a decline in

variety and research and development (Mueller & Tilton, 1969; Tushman & Anderson, 1986). The objective was to identify emerging or breakthrough inventions that were in the early stages of their S-curve and then to examine their thematic structure. However, due to the limited size of the dataset, the findings from the moderately small data on early patents had not yielded any significant conclusions. Consequently, the present study was constrained to an examination of CPC classes with the most substantial quantity. A thorough examination of the number of patents per class, subclass, group, and subgroup within the Y section was conducted (see Appendix 1 for the results of the complete analysis of quantities among the Y section). The two most prevalent items were selected for further analysis using the logistic S-curve to assess their position in relation to the curve.

Equation 1: *The logistic growth function.*

$$N(t) = \frac{K}{1 + exp[-r\,(t - t_0)]}$$

where:

- **N(t)** is the cumulative number of patents at time *t*.
- **K** is the *carrying capacity*, i.e., the upper saturation level of the curve (the theoretical maximum cumulative patents the technology will produce in its lifecycle).
- **r** is the *growth rate* (a positive constant that determines the steepness of the S-curve, reflecting how fast the innovation diffusion or patenting progresses),
- **t0** is the *inflection point*. This is the midpoint of the S-curve, marking the transition from accelerating growth to decelerating growth.

4.4.3 Latent Dirichlet Allocation (LDA)

This section discusses the application of Latent Dirichlet Allocation (LDA) to identify latent thematic structures within green energy storage patents. LDA, a generative probabilistic model, is particularly well-suited for identifying patterns in extensive unstructured text, such as patent abstracts and titles. The implementation is guided by established best practices in parameter tuning and model evaluation, with particular attention paid to coherence scoring as the principal measure of model quality. The selection of interpretable and semantically coherent topics is of particular importance, with these topics reflecting distinct technological subdomains.

In previous sections, we introduced text mining techniques that use machine learning to identify trends or patterns within patent documents. In this section, we will delve deeper into

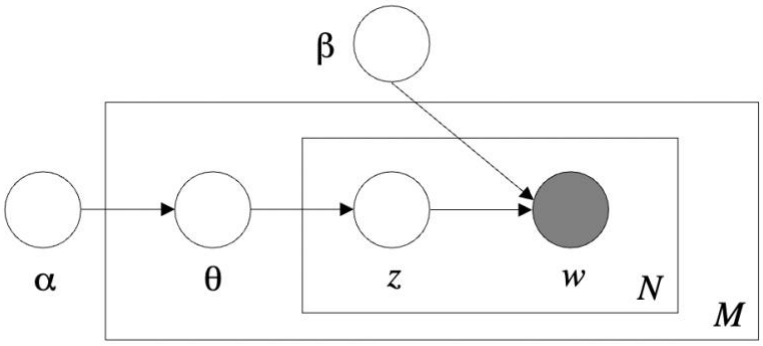
the fundamentals, exploring the underlying principles of LDA and the assumptions that were made to construct the model for this study.

Latent Dirichlet Allocation (LDA) was used for this study to identify hidden or latent themes from the green patent titles and abstracts. LDA is a probabilistic generative model that treats documents as a distribution over topics and topics as a distribution over words (Blei et al., 2003). Instead of delving into the formulas and equations that enable LDA, a graphical representation of how it works is presented in Figure 5 (Blei et al., 2003). The variables in the figure are as follows:

- **α (alpha)** = Prior knowledge of how mixed the topics are in the documents.
- **β (beta)** = Prior knowledge of how spread out the words are in the topics.
- **θ (theta)**: The specific topic mix for a given document.
- ***z* (z_n)** = The topic chosen for the nth word in the document.
- ***w* (w_n)** = The observed word itself.

The outer rectangle in the figure represents documents (patents), and for each document, we draw its topic mix θ (e.g., 15% Topic A, 70% Topic B, 15% Topic C). The inner rectangle represents words - for each word in the document, we (1) choose a topic z (which topic the word comes from) and then (2) we choose the actual word w from that topic (Blei et al., 2003).

Figure 5: *Graphical model representation of LDA.*



Source: Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, p-997.

Blei et al. (2003) refer to the important assumptions about LDA. The model posits that words are exchangeable and that the order of the words is irrelevant. This approach aligns with the *bag-of-words model*, which treats documents as a multiset of words (Blei et al., 2003). The authors further emphasize that *exchangeability* should not be confused with the assumption of independent and identically distributed (i.i.d.) observations. Instead, LDA posits that words are

conditionally i.i.d. given a latent topic distribution specific to each document. The assumption of *exchangeability*, or *bag-of-words*, posits that the meaning of a word is not contingent on its order in a document but rather on its frequency of occurrence (Gan & Qi, 2021). This text processing method significantly reduces the dimensionality of the represented text and is particularly advantageous when dealing with extensive text (Gan & Qi, 2021).

4.4.3.1 Parameter Selection and Model Training

The primary objective of this study is to identify latent substances within patent documents by employing an unsupervised LDA model. In this study, the *'LdaModel'* from the *'Gensim'* library was employed to execute LDA. The training of this model necessitated the establishment of several parameters, which will be the subject of the present discussion and justification.

The initial parameter pertains to the *number of topics* (*K*) that the model should generate. While higher values of *K* may capture more nuanced topics, they can also result in redundant or incoherent topics. To identify the most balanced approach, where topics are thematically distinguishable with as little overlap as possible, different numbers of *K* will be tested. Prior literature suggests that the standard number of topics is approximately 100, although this does vary depending on the specific dataset (Kaplan and Vakili, 2015; Blei et al., 2003). To this end, a topic range of 5-101 will be tested. The selection of *K* will be informed by validity tests, with a particular focus on coherence testing.

The subsequent parameter is the *number of passes*. This parameter defines how many full passes the model makes through the corpus during training. A higher number of passes improves convergence and stability of the learned distributions, particularly for medium to large corpora. The default setting for passes is 10, which will be utilized in this study (source).

The third critical parameter is alpha (α), which governs the distribution of topics across documents. In this study, an alpha was set to "auto," enabling the LDA model to learn an *asymmetric prior* from the data rather than relying on a fixed assumption. This enables the model to adapt the probability of topic occurrence in a document based on patterns identified in the corpus (source). Rather than assuming that all topics are equally likely to occur in all documents, this approach allows for variation in the occurrence of topics - some topics may occur frequently, while others may be more specialized or rare. The employment of an asymmetric, data-driven prior enhances the model's capacity to mirror real-world structures, in contrast to a symmetric prior that postulates equal likelihood for all topics irrespective of their content (source).

Finally, the fourth critical parameter, the *beta* (β), also called *eta* (η) parameter, was set to 0.01 (O'Callaghan et al. 2015), thereby establishing a *symmetric Dirichlet prior* over the distribution of words within each topic. This approach ensures that all words are initially treated equally, yet the low value of 0.01 enables each topic to be concentrated on a small number of highly relevant words, as opposed to the probability being spread over many terms. The strategic selection of a modest eta (η) facilitates the creation of focused and precise topics, a particularly advantageous approach when analyzing technical documents such as patents. In this context, the objective is to ensure that each topic reflects a specific theme or concept, such as "battery technology" or "solar panels," rather than incorporating a diverse array of unrelated terms. A sparse topic is more straightforward to interpret and offers a more meaningful perspective on the identification of patterns in the data. (source)

While the alpha parameter was set to "auto" to enable the model to learn the typical number of topics per document, the eta parameter was manually set to a small value. This was done to ensure that the topics themselves are clear, specific, and distinct. This methodological combination enhances the quality of the topic modeling and aligns with the objective of identifying meaningful thematic structures within the dataset.

4.4.3.2 Coherence

The quality of the LDA model utilized in this study was assessed by employing topic coherence as a pivotal evaluation metric. While traditional statistical measures, such as perplexity or log-likelihood, evaluate the probabilistic fit of a model, they often fail to correlate with the human interpretability of the topics (Chang et al., 2009). Conversely, topic coherence quantifies the semantic similarity between high-probability words within each topic, thereby offering a more interpretable and human-oriented evaluation of topic quality (Newman et al., 2010; Röder et al., 2015). Coherence is calculated by evaluating the degree to which words in a topic occur together in a reference corpus or similar contexts. In this study, the coherence measure 'c_v' available in the 'GenSim' library was employed.

The c_v measure integrates several steps to more accurately capture the semantic similarity between the top words in each topic. It combines four key components: segmentation, probability calculation, confirmation measure, and aggregation (Röder et al., 2015).

The first step in calculating c_v coherence involves taking the top *N* words from each topic (usually the top 10) and forming all possible word pairs; this is known as segmentation by a *set* or the *S-one-set method* (Röder et al., 2015). For example, if we take the words ["solar,"

"panel," and "energy"], we will create the pairs ("solar," "panel"), ("solar," "energy"), and ("panel, energy"). This way we avoid redundancy and reflect how humans naturally process related word sequences in language.

Next, the method builds up context vectors for each word in the pair using statistics of occurrence in the reference corpus. This is done by scanning the reference corpus with a 'sliding window,' usually about 110 words wide (Röder et al., 2015). Each word's context vector reflects how often it occurs near other words within these windows. The strength of co-occurrence between two words is quantified using Normalized Pointwise Mutual Information (NPMI), a statistical measure of association that captures how strongly two words are related relative to chance (Röder et al., 2015). NPMI is defined according to the following equation:

Equation 2: *Normalized Pointwise Mutual Information (NPMI).*

$$NPMI(wi, wj) = \frac{log\ (\frac{P(wi, wj)}{P(wi)\ x\ P(wj)}}{-log\ P(wi, wj)}$$

where:

- P(wi) is the probability that the word wi appears in the sliding window
- P(wi,wj) is the probability that both words occur simultaneously in the same window

The NPMI score ranges from -1 (words never co-occur) to 1 (words always co-occur), where 0 means that the words are statistically independent. Once we have a vector representation of each word (based on its NPMI values relative to all other words), we calculate the cosine similarity between each pair of word vectors (Röder et al., 2015). The cosine similarity measures how aligned the two vectors are, regardless of their length. If two words frequently occur in similar contexts, their vectors will point in similar directions, resulting in a high cosine similarity score close to 1. Mathematically, cosine similarity is defined as follows:

Equation 3: *Cosine similarity.*

$$cos(vi, vj) = \frac{vi \cdot vj}{||\ vi\ ||\times||\ vj\ ||}$$

where:

- *vi* and *vj* are two vectors (e.g., topic distributions),
- ||*vi*|| and ||*vj*|| are the magnitudes (lengths) of the vectors.

Resulting value ranges from:

- +1: the same direction (highly similar),
- +0: orthogonal (no similarity),
- -1: opposite directions (rare in topic models).

This step ensures that we are not just counting how often words co-occur but rather evaluating whether they share similar semantic environments in natural language.

Finally, the mean of all pairwise cosine similarities is computed to obtain a single coherence score for the topic (Röder et al., 2015). The resulting value reflects the overall semantic coherence of the top words of the topic. Higher coherence values indicate that the topic words are not only frequent but also conceptually related in natural language use.

Equation 4: *Coherence Score.*

$$Coherence_{cv} = Avarage\{(v_{wi}, v_{wj}) \; for \; all \; (wi, wj)\}$$

This process is repeated for each subject in the model. The coherence score for all topics can then be averaged to give an overall measure of the model's quality (Röder et al., 2015).

The c_v measure was selected for its demonstrated capacity to correlate highly with human judgments across multiple datasets and settings. In their study (Röder et al., 2015), the researchers compared different coherence measures. They found that the c_v measure performed best on average across different data sources (tested 237,912 coherences), and the measure most closely correlated with human judgments. The selection of coherence as a metric is further substantiated by prior empirical research indicating that coherence-based evaluations surpass perplexity-based evaluations in identifying interpretable topic structures (O'Callaghan et al., 2015).

The coherence scores were calculated for models with varying numbers of topics (k), and the average coherence was used to guide model selection. Following established best practices, coherence was employed not only to ascertain the optimal number of subjects but also to validate subject quality post hoc (Stevens et al., 2012). Higher coherence scores indicate more semantically consistent groupings of topic words, which is desirable for interpretability and downstream analysis.

4.4.3.3 Topic Labeling

Following the execution of LDA modeling and the analysis of the ten most frequent keywords, it is essential that the topics are assigned an overall title that effectively encapsulates the central theme. In this study, *ChatGPT* was used to name the identified topics created with LDA. Previous research has used methods such as domain experts for identification and naming (Chuang et al., 2012). However, given my limited experience in engineering and technical terms and limited time to reach out to experts, *ChatGPT* was used to label the topics. This was done by assigning the AI the ten most frequent words within each topic as a basis for

latent topic labeling. The prompt used to generate the topic labels from ChatGPT can be reviewed in Table 11. While AI-based topic labeling provides efficiency, it may introduce interpretive bias or lack domain-specific nuance compared to expert validation. Furthermore, a personal evaluation of the labels was conducted to assess their suitability. This was not done out of blind faith in generative AI, but rather to ensure, through critical review, that the labeling represents the keywords.

Table 11*: Prompt Used for AI-Assisted Topic Labeling Based on Top Keywords from LDA Output.*

"For each topic generated by the LDA model, examine the top 10 keywords associated with that topic:
[topic 1: keyword 1-10, topic 2; keyword 1-10,....., topic n: keyword 1-10]
Based on these keywords, assign a concise and descriptive human-readable label that captures the underlying technical or conceptual theme of the topic.
The label should:

- *Reflect the dominant concept or mechanism implied by the keywords.*
- *Use technical terminology consistent with the domain.*
- *Avoid redundancy or vague terms like "system" or "component" unless uniquely meaningful.*

Note: *This table presents the prompt provided to ChatGPT for the purpose of assigning descriptive and technically relevant labels to topics generated through Latent Dirichlet Allocation (LDA). Each topic was represented by its ten most probable keywords. Given the author's limited expertise in engineering terminology and time constraints preventing expert consultation, an AI-based labeling strategy was employed. The prompt instructed ChatGPT to generate concise, domain-appropriate topic labels based on the top keywords, avoiding generic or ambiguous terms.*

4.4.4 Patent Topics Visualization

This section presents the visual tools used to interpret the latent topics uncovered by the LDA model applied to patent data. Two complementary approaches, LDAvis and t-distributed Stochastic Neighbor Embedding (t-SNE), are employed to enhance the interpretability of topic structures and document-level relationships. LDAvis provides a comprehensive analysis of topic distributions, intertopic distances, and the most informative terms per topic, enabling precise semantic interpretation. Concurrently, t-SNE offers a low-dimensional spatial projection of each patent's topic composition, thereby enabling the identification of coherent thematic clusters and outliers. Collectively, these tools enhance the transparency and reliability of the topic modeling process, facilitating both qualitative understanding and subsequent quantitative validation.

4.4.4.1 Intertopic Distance Map—LDAvis

In this study, an interactive visualization tool called LDAvis is used to better understand, interpret, and evaluate the latent topics generated by LDA (Sievert & Shirley, 2014). An intertopic distance map and a ranking of the most relevant terms for each topic are two visual components provided by the Python library LDAvis that enhance the interpretability of topic models. These elements make it easier to examine the semantics of individual topics as well as to get a general picture of topic relationships (Sievert & Shirley, 2014).

The intertopic distance map is a two-dimensional representation of the relationships between topics. Multidimensional scaling (MDS) is used to reduce dimensionality after LDAvis uses a divergence measure to estimate the pairwise distances between topic distributions to compute this visualization (Sievert & Shirley, 2014). The result is a scatter plot where each topic is visualized as a circle. The size of the circle represents the total occurrence of the topic in the corpus, while the distance between the circles reflects the topic similarity. Topics that are closer together are assumed to have more similar term distributions, making it easier to visually identify topic clusters and outliers (Sievert & Shirley, 2014).

The second component that LDAvis contributes is an overview of the 30 most informative terms within each topic. Traditional topic exploration methods typically list the most likely words within a topic, but this approach often highlights overly general terms that lack distinctiveness. To remedy this, Chuang et al. (2012) introduced two complementary measures: *distinctiveness* and *saliency*.

Distinctiveness is defined as the deviation between the conditional probability of a topic and the marginal topic distribution across the corpus. Terms that occur in many topics are considered to be less distinctive, while terms that are concentrated in a small number of topics are considered to be more distinctive (Chuang et al., 2012).

To balance term frequency and topic specificity, the authors define *salience* as the product of a term's probability in the corpus and its distinctiveness. This composite measure identifies terms that are both frequent enough to be meaningful and specific enough to be informative for topic interpretation (Chuang et al., 2012).

Building on these ideas, LDAvis implements a flexible method for ranking subject terms according to a metric called *relevance* (Sievert & Shirley, 2014). *Relevance* allows users to set the balance between term frequency within a subject and distinctiveness at the corpus level using a parameter λ. When $\lambda=1$, terms are ranked only by their topic-specific probability. The optimal value, determined through user studies, is $\lambda=0.6$, which balances specificity and interpretability (Sievert & Shirley, 2014). This relevance-based ranking is presented in a bar

graph showing both the subject-specific frequency (in red) and the corpus-wide frequency (in gray) of each term. This allows users to quickly infer whether a term is informative due to its frequency, specificity, or both (Sievert & Shirley, 2014). Formally, the *relevance* of a term (w) to a topic (k) given a weight parameter (where $0 \leq \lambda \leq 1$) is defined as follows:

Equation 5: *Definition of Relevance.*

$$r(w, k \mid \lambda) = \lambda log(\phi_{kw}) + (1 - \lambda) log(\frac{\phi_{kw}}{p_w})$$

where ϕ_{kw} is the probability of word w in topic k, p_w is the marginal probability of w in the corpus, and λ is a tunable parameter. A user study conducted by Sievert and Shirley (2014) found that a value of λ = 0.6 provides an optimal balance between specificity and informativeness, outperforming traditional rankings based on probability alone.

4.4.4.2 t-distributed Stochastic Neighbor Embedding (t-SNE)

To help explore and understand patterns in the text of green patents, this study uses a technique called *t-distributed Stochastic Neighbor Embedding* (t-SNE). This method is especially useful for visualizing complex, high-dimensional data, such as the kind produced by topic modeling with LDA, where each document is represented as a mixture of topics (Blei et al., 2003). Because these representations often have dozens or even hundreds of dimensions, it becomes difficult to directly interpret or visualize them. t-SNE helps by projecting this high-dimensional data into a low-dimensional space, typically two or three dimensions, so that it can be plotted and visually analyzed (Blei et al., 2003).

Van der Maaten and Hinton (2008) presented t-SNE as an evolution of previous methods (stochastic neighbor embedding - SNE) for visualizing high-dimensional data. The basic idea behind t-SNE is to measure how similar different documents are in their original (high-dimensional) form and then try to place them in a low-dimensional space in such a way that these similarities are preserved. To do this, t-SNE calculates how likely it is that two documents are "neighbors" (i.e., similar to each other) and then tries to ensure that documents that are close to each other in the original space remain close to each other in the final visualization. To this end, a probability distribution is used to help highlight clusters or groups of documents that have similar topic patterns (Van der Maaten and Hinton, 2008).

A challenge with t-SNE is that its results are contingent on several parameters, such as perplexity, learning rate, and exaggeration, which control how the algorithm balances different patterns in the data. More recent research by Gove et al. (2022) has offered new, more efficient default settings for these parameters. They also introduced an automated method for selecting

hyperparameters based on the characteristics of the dataset, which helps users avoid guesswork and improve the quality of their visualizations. Their results suggest that t-SNE performs best when using relatively small values of perplexity and a learning rate that scales with the number of data points.

In this thesis, t-SNE is used to visualize the subject distribution of green patents according to the LDA model. The visualization of each patent in a two-dimensional space is based on its subject composition, facilitating the emergence of patent clusters with similar themes. This approach enables the identification of patterns in green innovation, such as groups of patents focusing on renewable energy, waste management, or sustainable materials. Furthermore, the visualization assists in the identification of unusual patents that combine topics in novel ways or diverge from established trends.

To visualize the patent subject distributions (θ-matrix) from the LDA model, the t-distributed stochastic neighborhood embedding (t-SNE) algorithm with multiple configurations was used to test robustness and cluster interpretability. Rather than selecting arbitrary parameters, three distinct configurations (A, B, and C) were assessed to ascertain the most effective representation of subject structures in two dimensions (refer to Table 12).

Table 12: *The Different Configurations of Parameters Tested in the t-SNE Algorithm.*

Config.	Perplexity	Early Exaggeration	Learning Rate	Description
A	30	12.0	auto	Baseline (Default)
B	40	16.0	200	More global separation
C	20	12.0	100	More local separation

Note: *The table summarizes three t-SNE configurations tested to visualize topic structures. Configuration A uses default settings, B enhances global separation, and C emphasizes local detail. These variations help assess topic clustering stability and interpretability.*

For each configuration, the algorithm was initialized with PCA (init='pca') to enhance stability and convergence (scikit-learn, n.d.). The Barnes-Hut approximation method was employed (method='barnes_hut') with an angle=0.5 to balance precision and performance, given the moderate size of the dataset (scikit-learn, n.d.). A maximum of 1000 iterations was permitted for the purpose of achieving convergence (scikit-learn, n.d.). To assess which configuration offered the most coherent topic grouping, k-means clustering was applied to the resulting 2D t-SNE coordinates, followed by computing the silhouette score (see more details about the silhouette score in section 4.4.4.2.1).

4.4.4.2.1 Silhouette Score

The silhouette score, introduced by Rousseeuw (1987), is a widely utilized metric for evaluating the quality of cluster assignments. The silhouette score provides a quantitative measure of how well an individual item fits within its assigned cluster in relation to other clusters (Rousseeuw, 1987). In other words, it quantifies the degree to which patent documents align with the topic keywords generated by LDA. This measure plays a central role in both the interpretation of the resulting cluster structure and the validation of the choice of the number of clusters. The silhouette score for object *i* is defined as follows:

Equation 6: *The Silhouette Score.*

$$s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}$$

where:

a(i) is the average dissimilarity between *i* and all other objects in the same cluster, and *b(i)* is the minimum average dissimilarity between *i* and all objects in any other cluster (Rousseeuw, 1987).

The value of $s(i)$ ranges from -1 to +1. When values are close to +1, it indicates that the object is well aligned with its own cluster and poorly aligned with neighboring clusters. When values are around 0, it indicates that the object lies between two clusters and can be classified in an ambiguous way. Finally, when values approach -1, it indicates that the object is likely to be misclassified and would be better assigned to another cluster (Rousseeuw, 1987).

4.5. Limitations

This study employs scalable, data-driven techniques to explore green innovation within green patents; however, several methodological constraints must be acknowledged. These include dataset restrictions (limited export volume and language bias), model-based assumptions (e.g., bag-of-words), absence of expert validation, and simplifications inherent in topic modeling and maturity assessment frameworks. Despite the implementation of rigorous methodologies, including coherence scoring, standardized preprocessing, and logistic modeling, these limitations introduce potential biases in representation, interpretation, and generalizability of the findings.

Although this study applies robust, data-driven methods to analyze green innovation patents, several limitations must be acknowledged in order to provide a balanced account of its methodological rigor. The dataset was obtained from Lens.org, an open-access platform that aggregates patent data from major international sources, including the EPO, USPTO, and WIPO. However, due to inherent limitations, the maximum number of patents that could be exported was 50,000. This limitation introduces a risk of sample bias and limits the external validity of the results (Bryman & Bell, 2015). As the selection mechanism for these records is undocumented and assumed to be random, it is uncertain whether the dataset adequately reflects all relevant green patent developments globally.

To ensure data consistency, the text corpus was restricted to English-language content, and non-ASCII characters were removed. This reduced the complexity of the pre-processing stage. However, it also resulted in the exclusion of patents from key innovation ecosystems, such as those found in China, Japan, and Korea, which use their local language in patent applications. This may potentially skew the results in Western contexts (Magerman et al., 2015).

LDA assumes a bag-of-words model that ignores word order, grammar, and contextual nuances. Although suitable for large-scale text analysis, this simplification may limit the semantic precision of topics (Blei et al., 2003). Although metrics of topic coherence and silhouette scores were used to assess quality, the lack of validation by domain experts' limits face validity and may compromise the conceptual accuracy of extracted themes (Chang et al., 2009; Chuang et al., 2012).

The logistic S-curve was employed as a metric to evaluate technological maturity, with the assessment based on patent volume over time. While this approach has a robust foundation in the innovation management literature (Gao et al., 2013; Haupt et al., 2007), it presupposes a predictable growth-maturity-saturation trajectory. However, in practice, innovation trajectories frequently exhibit non-linear patterns and are influenced by sudden regulatory shifts, market volatility, or the emergence of disruptive technologies (Tushman & Anderson, 1986). Furthermore, the utilization of the number of patents as a proxy for maturity can be misleading. This is due to the fact that strategic or defensive patenting has the capacity to inflate activity without reflecting real progress (Hall et al., 2005).

The use of generative AI (ChatGPT) to label LDA topics improved efficiency and consistency, but it also brought a layer of automation bias. Unlike expert panels, AI lacks domain knowledge to interpret technical nuances, which can affect the objectivity and accuracy of topic names. Although steps were taken to mitigate the bias through the use of top keywords

and thematic coherence, this trade-off between speed and depth of interpretation remains a limitation of the study.

4.6 Research Quality Criteria

This section evaluates the methodological rigor of the study through established research quality dimensions: reliability, validity, objectivity, and replicability. Drawing on guidelines from Bryman and Bell (2015), the purpose is to critically assess how well the data collection, preprocessing, modeling, and interpretation procedures meet the standards for credible and trustworthy research. Particular attention is given to the use of algorithmic tools, coherence-based evaluation, and transparent workflows, as well as acknowledged limitations such as the lack of expert-coded validation and linguistic bias. This appraisal strengthens the academic foundation of the study and frames the interpretive confidence of its findings.

4.6.1 Reliability

Reliability is defined as a consistent and stable measurement process (Bryman & Bell, 2015). In this study, reliability was achieved through the standardization of Python-based libraries (e.g., Gensim, scikit-learn), the fixing of random seeds for topic modeling, and the implementation of a systematic pre-processing procedure. The application of these procedures serves to minimize variance due to contributions from random calculations, thereby ensuring a high degree of internal reliability (Bryman & Bell, 2015).

4.6.2 Validity

There are several dimensions of validity analysis (Bryman & Bell, 2015). Construct validity encompasses the aspect of the patents under the Y02/Y04 CPC system, which is then matched with already established classifications for climate change mitigation technologies (Favot et al., 2023) to ensure conceptual alignment. To ensure face and content validity, the measures are to consist of coherence measures as well as keyword inspection. However, the absence of domain expert review is a limitation that reduces the strength of face validity (Chang et al., 2009). The external validity of the dataset is constrained by the quantity of documents contained therein; furthermore, the exclusion of non-English documents serves to limit the generalizability of the results to Western regions alone (Bryman & Bell, 2015).

4.6.3 Objectivity

The objective is to minimize researcher bias through the implementation of algorithmic modeling with open-access tools and transparent pre-processing workflows. Nonetheless, the efficacy of AI-assisted topic labeling is evident; however, it is imperative to acknowledge the inherent interpretative bias that is embedded within the process, given that the documents are not subject to the codings of experts within the technical domain. This aspect is widely acknowledged as a means to enhance methodological transparency.

4.6.4 Replicability

All tools, methods, and datasets applied in this study are either openly available (e.g., Lens.org) or reproducible through widely adopted Python libraries. The detailed pre-processing, model parameters, and all evaluation metrics allow any other researcher to replicate or even extend the study in different contexts (Bryman & Bell, 2015).

5. Result

This chapter presents the main empirical results derived from patent trend analysis and LDA topic modeling. It begins with a growth analysis of CPC categories using logistic S-curves and narrows the focus to the selected subclass Y02E60/10. The chapter then presents the topics discovered through LDA, followed by visualizations using LDAvis and t-SNE to validate topic structure and relationships. These results provide a detailed semantic landscape of innovation activity in energy storage technologies.

5.1 Patent Trend Analysis and Subgroup Selection.

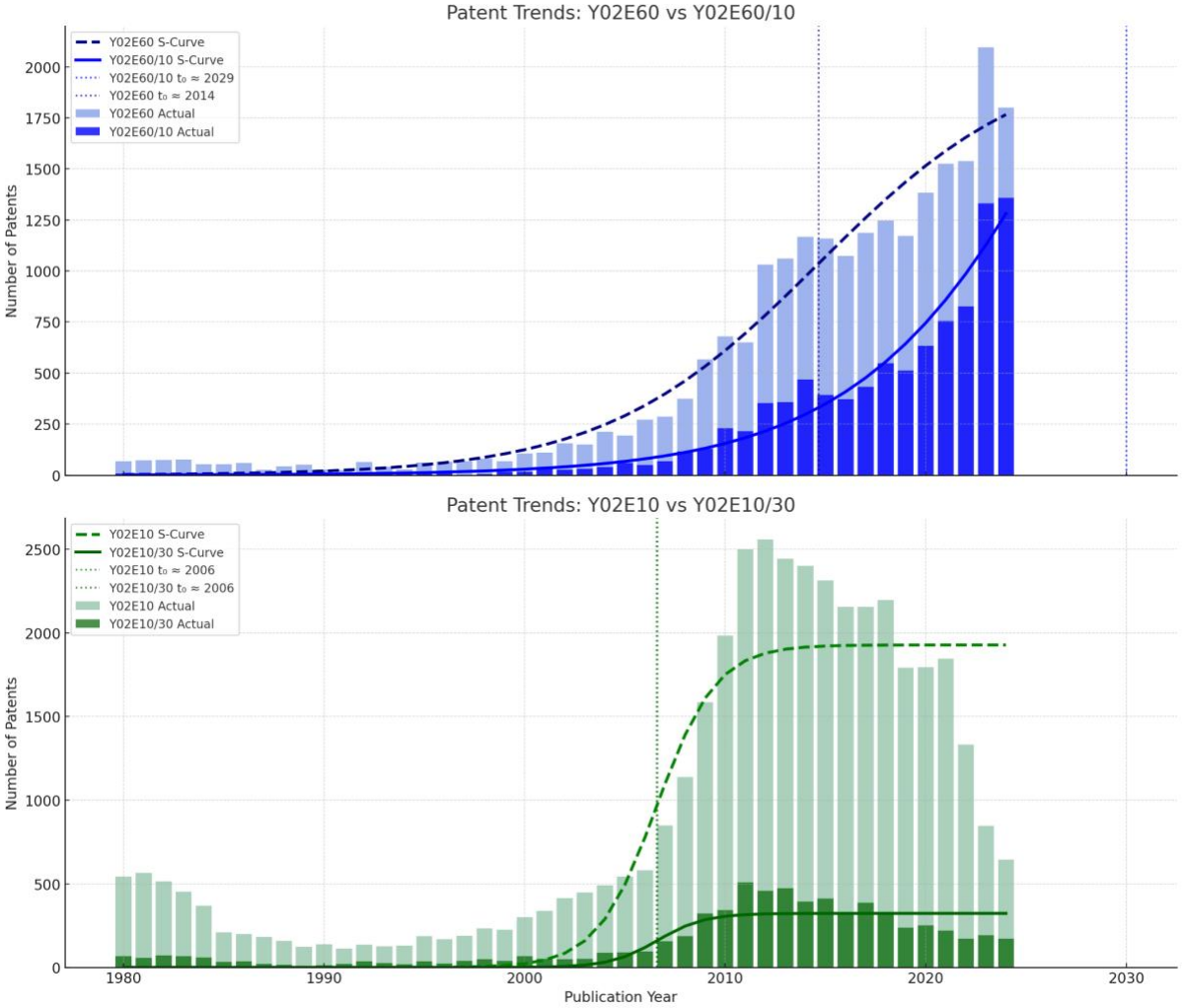
To identify the most promising CPC subgroup for topic modeling, patent trend analysis was conducted using S-curve modeling across several Y02E subgroups. The emphasis of this study will be on the two most prevalent groups and their most prevalent subgroup within the dataset. This analysis was crucial to identify innovation maturity levels and forecast future technological development trajectories. The complete trend analysis resulting in the most frequent classifications can be found in Appendix 1. The selection of the two most frequent CPC groups (Y02E10 and Y02E60) and their most frequent subgroups (Y02E10/30 and Y02E60/10) for further analysis was predicated upon the analysis outlined in Appendix 1.

As presented in Figure 6, the top panel displays a comparative S-curve and the actual trend of the number of patents between Y02E60 (general energy storage technologies) and its

subgroup Y02E60/10 (electrochemical energy storage systems, including batteries and related innovations). Y02E60/10 demonstrates a marked increase in patent applications from 2015 onward, indicative of a period characterized by substantial innovation and commercialization. According to the S-curve forecast, this subgroup will reach its inflection point around 2024, with an estimated maturity around 2030. Conversely, the Y02E60 group (parent group representing a more extensive array of technologies) attained its inflection point at a considerably earlier point, around 2014, and is currently entering a maturity plateau. The data suggest that Y02E60/10 is still in the growth phase of its technological lifecycle, making it an ideal candidate for LDA topic modeling. The technology is dynamic and evolving; therefore, topic mining can provide insights into emerging themes, technological challenges, and innovation patterns that have not yet been consolidated into a mature knowledge domain.

The bottom panel of Figure 6 illustrates Y02E10 (renewable energy production) and its subgroup Y02E10/30 (hydrogen-related technologies). A decline in patent activity has been observed in both groups in recent years. The S-curve for Y02E10/30 indicates that its inflection point occurred prior to 2010 and that the growth phase concluded around 2015-2018. This decline in application activity suggests that the field may be approaching technological saturation or has entered a period of incremental innovation. Given the apparent decline in novelty and exploration within Y02E10/30, the application of LDA would likely yield more repetitive and saturated topics, thereby limiting its strategic value.

Figure 6: *Comparative Patent Trend Analysis Using S-Curve Modeling of Two CPC Key Groups and Subgroups.*



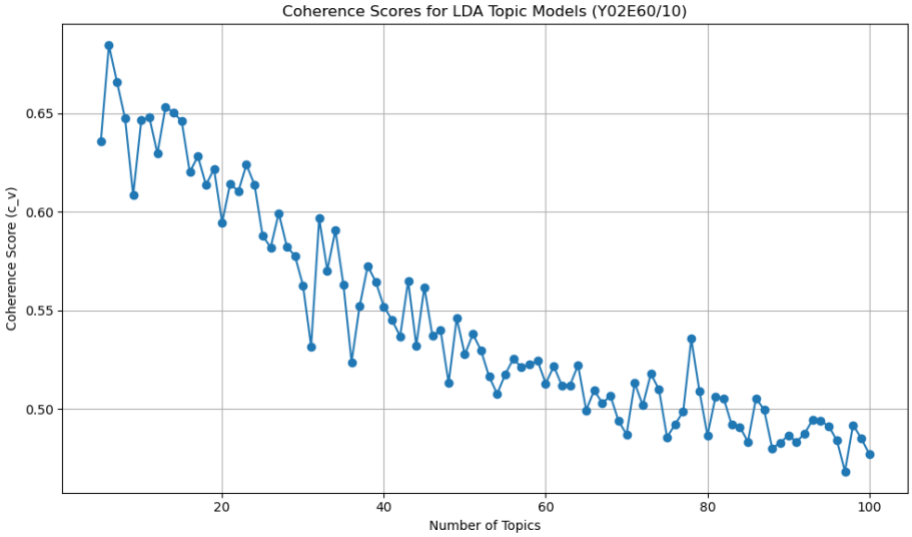
Note: *Patent trends and S-curve diffusion models for CPC subgroups Y02E60/10 and Y02E10/30. The upper panel compares overall energy storage technologies (Y02E60) with the electrochemical storage subgroup (Y02E60/10). The lower panel compares general renewable energy generation (Y02E10) with hydrogen-related systems (Y02E10/30). Dashed lines represent logistic growth forecasts; vertical dotted lines mark estimated inflection points and estimated points of maturity.*

In light of these observations, the remaining analysis will concentrate exclusively on the subgroup Y02E60/10, which pertains to the domain of energy storage utilizing batteries. The dataset under consideration contains 7,211 patents that fall into this classification. As stated in the methodology chapter, the dataset did not exclusively comprise patents containing English descriptions, thereby necessitating an examination of the impact of language noise. The analysis revealed that 425 patents out of 7,211 (5.89% of the filtered Y02E60/10 dataset) contained non-English content. These patents were not removed, but their non-ASCII words were excluded from the topic modeling. Consequently, 94.1% of the model is predicated on purely technical terms in English only.

5.2 Selection of Number of Topics.

The next phase of the patent analysis, employing CPC code Y03E60/10, entailed the determination of the optimal number of topics to utilize for the LDA modeling. As stated in the methodology, a range of 5 to 101 topics was initially assessed. This range was subsequently evaluated using the coherence score to ascertain the optimal number of topics that would contribute to the optimal topic interpretability. Figure 7 illustrates the coherence score for each number of topics. The graph indicates that the LDA model with six topics (coherence score = 0.6945) achieves the highest score, suggesting that this number of topics is optimal for the dataset. The graph shows a clear decrease in coherence as more topics are created from the dataset. The lowest observed coherence score was 0.4683 and was found at 97 topics.

Figure 7: *Coherence Score Test Results.*



Note: *Coherence scores (c_v) for LDA models trained on Y02E60/10 patent abstracts and titles. The c_v coherence metric evaluates semantic similarity among topic keywords. Results indicate optimal topic interpretability at 6 topics (out of a range of 5-101).*

5.3 LDA Topic Modeling Results.

This section presents the empirical results of the application of LDA to a dataset of patents within the CPC subgroup Y02E60/10, which relates to technologies related to energy storage using batteries. The LDA model was trained on the titles and abstracts of these patents, using pre-processed text data that included bigram and domain-specific stopword removal. A primary objective of this analysis was to identify latent thematic structures that reflect technological innovation patterns. The objective was to ascertain the optimal number of topics (K) by calculating coherence scores over a range of topic models. As demonstrated in the previous section, the highest average coherence was achieved with six topics, and then

coherence steadily decreased (Figure 7). Consequently, a 6-subject model was selected as the most robust, exhibiting an optimal balance between coherence and thematic granularity.

Table 13 offers a summary of the top 10 keywords for each topic and the corresponding labels assigned by ChatGPT. It also provides a rationale for the label, which aims to promote transparency regarding the underlying logic of the generative AI approach. Each topic represents a distinct technological focus within the broader scope of battery systems. It is noteworthy that the topics encompass thermal engineering (Topic 1), mechanical design (Topic 2), electronics and circuitry (Topic 3), physical housing and structure (Topic 4), electrochemistry (Topic 5), and safety/modularity concerns (Topic 6).

Table 13: *LDA-Derived Thematic Topics, Top Keywords, and Interpretive Labels for Y02E60/10*

Topic	Top 10 Keywords	Assigned Label	Rationale Behind Label
Topic 1	cooling, cells, liquid, heat, temperature, liquid_cooling, group, packs, water, heat_exchange	**Thermal Management in Battery Packs**	"Captures cooling, temperature, and heat exchange in liquid-based systems for battery groups."
Topic 2	shell, heat_dissipation, frame, mounting, air, container, fixing, formed, supporting, equipment	**Battery Enclosure & Heat Dissipation**	"Focused on the mechanical frame, shell design, and thermal release via air structures."
Topic 3	electric, voltage, charging, management, circuit, current, temperature, output, vehicle, cells	**Electric Control & Charging Circuits**	"Encompasses energy flow regulation, voltage/current control, and EV-related charge systems."
Topic 4	portion, terminal, surface, cells, housing, cover, case, direction, member, having	**Terminal Housing & Interface Structure**	"Suggests physical terminal components and structural aspects of cell casing and interfaces."
Topic 5	layer, material, electrolyte, electrochemical, comprising, positive, negative, active_material, anode, cathode	**Electrochemical Layer Materials**	"Core electrochemistry and materials science: electrodes, electrolytes, layers in cell makeup."
Topic 6	housing, store, layer, stack, comprising, board, contact, region, protection, new	**Battery Stack & Protection Modules**	"Suggests stacked housing units, contact boards, and layered protection systems in batteries".

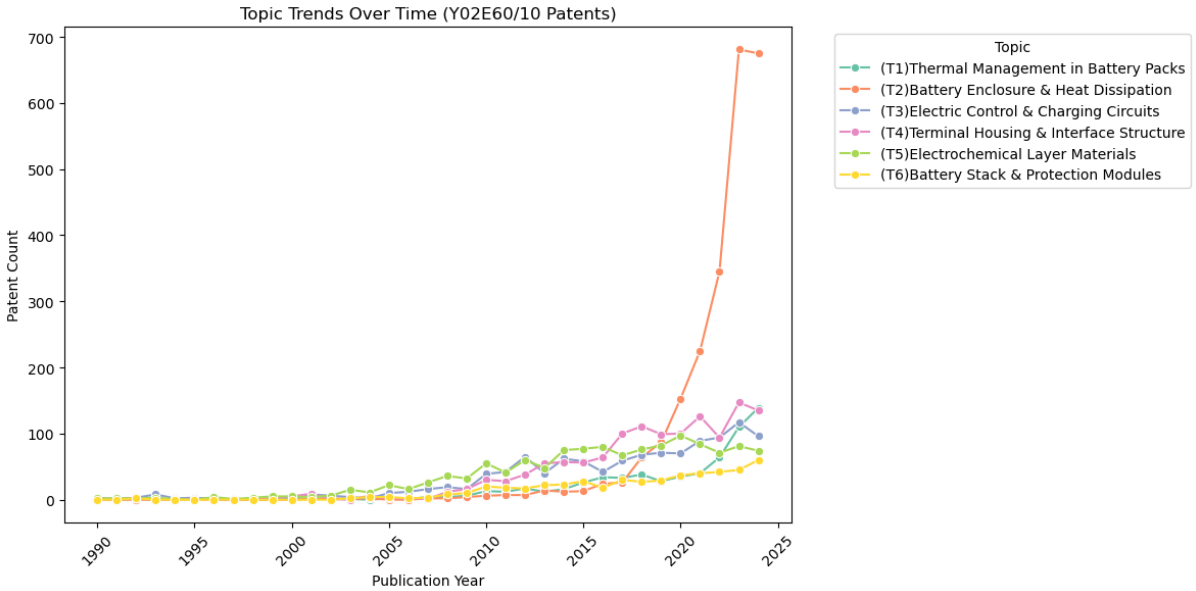
Note: *This table summarizes the six topics identified through LDA applied to a filtered corpus of patents within the Y02E60/10 subgroup. Each topic was assigned a readable label based on keyword content and domain relevance. The rationale column offers transparent justification for the label, produced with generative AI support.*

Figure 8 illustrates the temporal progression of six latent subjects identified within the CPC subclass Y02E60/10. The figure illustrates how the occurrence of each topic has changed

over time based on the annual number of patents associated with each theme from 1990 to 2024. The most notable trend is observed in T2 (battery enclosures and heat dissipation), which demonstrates an exponential growth in patent activity from around 2017 onwards. By 2023, the number of patents in this subject had increased to almost 700, indicating a substantial acceleration in technological development. Additionally, Topic 3 (Electrical control and charging circuits) has demonstrated a consistent upward trajectory, particularly after the year 2010. Similarly, Topic 4 (Terminal housing and interface structure) demonstrates consistent growth, reaching a peak shortly after 2020. In contrast, topic 5 (electrochemical layer materials) demonstrates a more gradual and steady increase in the number of patents over the years, followed by a leveling off after 2020. This finding suggests the presence of ongoing yet stable research activity in the development of electrode and electrolyte materials, which may be indicative of the maturation of specific sub-areas of battery chemistry. T1 (heat management in battery packs) demonstrates moderate growth, with an accelerating increase after 2020. However, it remains less dominant than the subject of battery chemistry. Finally, topic 6 (battery stacks and protection modules) has demonstrated the least frequent level of patent activity, although it has exhibited a gradual yet consistent growth trajectory.

The figure indicates a marked shift in recent years toward system-level issues, particularly those related to the thermal and structural aspects of battery packs. This shift can be attributed to the industry's response to the challenges of scaling up and implementing high-density battery technology in demanding applications.

Figure 8: *Topic trends over time for patents classified under Y02E60/10.*



Note: *Topic trends over time for patents categorized under Y02E60/10 (battery innovation technologies). Each line represents the annual frequency of patents dominated by a given LDA topic. Notably, Topic 2 (Battery Enclosure & Heat Dissipation) exhibits an exponential rise in recent years, signifying a shift in innovation toward*

mechanical and thermal optimization of battery structures. The (T1,…,T6) markings preceding the label refer to the specific topic numbers created with LDA to facilitate comprehension.

5.4 Intertopic Distance Map and Most Relevant Terms.

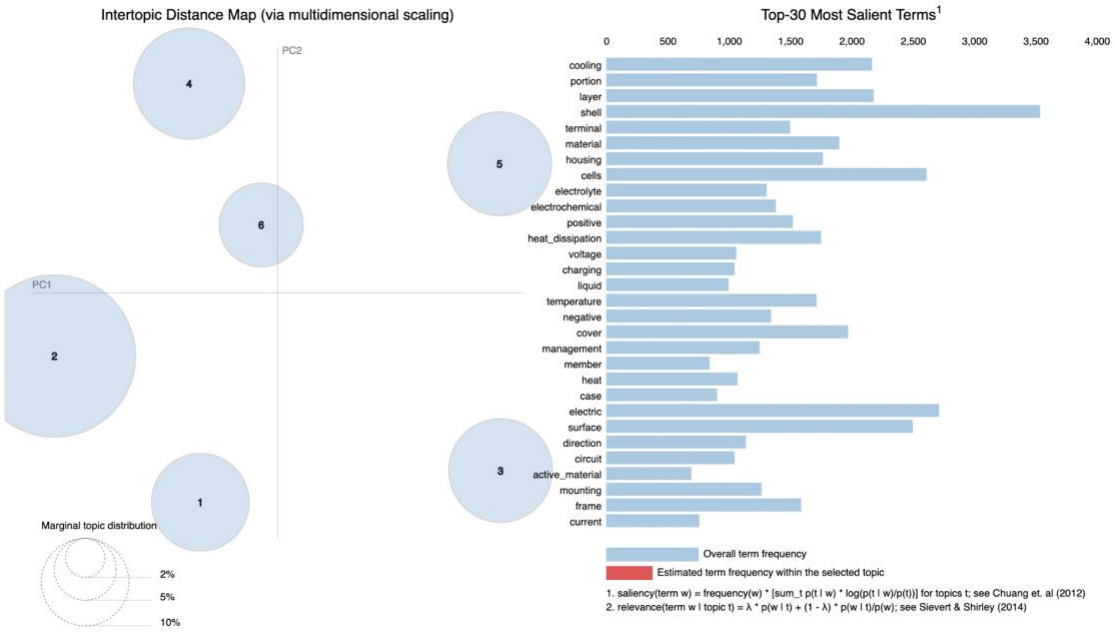
To better understand the semantic relationships between topics and to further validate the interpretability of the LDA model, *gensimvis* from the Python library *pyLDAvis* was used to visualize both distances between topics using multidimensional scaling and illustrate the 30 most relevant terms per topic, according to the relevance measure $\lambda = 0.6$. This tool provides an interactive and intuitive view of how well separated the topics are and what characterizes each topic in terms of its high-probability terms.

On the left side of the pyLDAvis output (Figure 9), an intertopic distance map, generated via multidimensional scaling (MDS), is displayed. Each bubble represents a topic, and the size of the bubble corresponds to how common the topic is in the whole corpus. The distance between the bubbles reflects semantic distinctiveness: closely located topics share overlapping vocabulary, while distant bubbles represent more unique thematic content.

The result with the LDA model with six topics for the Y02E60/10 patents is shown in Figure 9. Topic 2 is the most prominent topic in the corpus, with a significantly larger size relative to the other topics and a substantial coverage of the marginal topic distribution in the bottom left portion of the graph. This finding is indicative of its substantial presence in the documents. Topics 3 and 5 are located on the right side of the map, with moderate size and clear separation from other topics. The placement of these elements suggests a strong semantic uniqueness, which is likely indicative of distinct technical themes, such as electrochemical materials (T5) or electrical control systems (T3). T1 is situated within the lower left quadrant, positioned below T2 and to its immediate right. Its moderate size and proximity to T2 may indicate some overlap or adjacent relevance, possibly related to applications or thermal systems, as determined by previous keyword interpretation. T4 is situated in the upper left quadrant, a considerable distance from T3 and T5, while being in closer proximity to T6. This configuration may be indicative of a limited thematic association, such as structural design, terminals, or housing, with minimal overlap. T6 occupies a central position in the plot, exhibiting a small to medium circle size. This finding indicates that T6 is semantically average or general, potentially overlapping with elements from other themes but without being dominant. This finding aligns with the potential function of the subject as a bridging theme, as evidenced by its association with storage modules and generic protection functions. This visual

mapping confirms that the chosen number of topics provides a balance between granularity and separation, with each topic occupying a distinguishable part of the semantic space.

Figure 9: *Intertopic Distance Map and the Top 30 Most Salient Terms.*



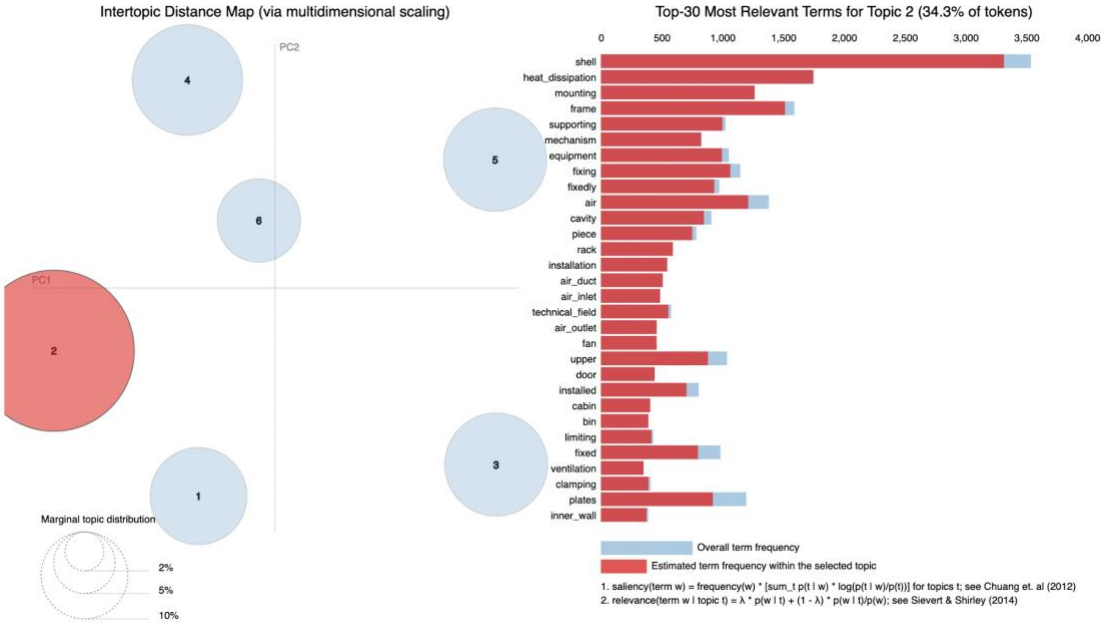
Note: *The intertopic distance map visualizes the semantic relationships among the six topics derived from LDA modeling of Y02E60/10 patents. Each circle represents a topic, with its size indicating prevalence in the corpus and its position reflecting relative semantic distance via multidimensional scaling. The minimal overlap between circles suggests well-separated and coherent topics, confirming the interpretability and thematic distinctiveness of the selected model.*

The right panel of the *pyLDAvis* interface presents the 30 most relevant terms associated with each latent topic generated by the LDA model. These terms were visualized using a relevance parameter ($\lambda = 0.6$), a setting recommended by Sievert and Shirley (2014) to enhance the interpretability of topics by prioritizing terms that are not only frequent within a topic but also relatively distinct in comparison to the entire corpus. Each bar in the visualization is represented by a two-color format: red bars signify the estimated term frequency within a specific topic, while blue bars denote the overall frequency of the term across the entire corpus. This format effectively separates subject-specific terminology from generally frequent terms, facilitating accurate and domain-relevant analysis.

Figure 10 illustrates an example of how to interactively select between topics and study topic-specific terms. The example in Figure 10 focuses on Topic 2, which represents the largest part of the corpus. A list of the top terms within T2 includes terms such as "shell," "heat dissipation," "assembly," "frame," "support," "mechanism," "fixing," "air duct," and "rack." This topic focuses on the mechanical and structural enclosure of batteries, including casings and ventilation solutions. A preliminary analysis of key terms indicates a focus on innovations in battery enclosures, mounting frames, and cooling via air circulation. The analysis reveals

that this particular topic occupies a predominant position within the corpus, accounting for 34.3% of all tokens. In other words, it can be deduced that 34.3% of all words contained within the dataset's patent titles and abstracts are likely derived from T2, thereby underscoring its pivotal role in the conceptualization of battery systems.

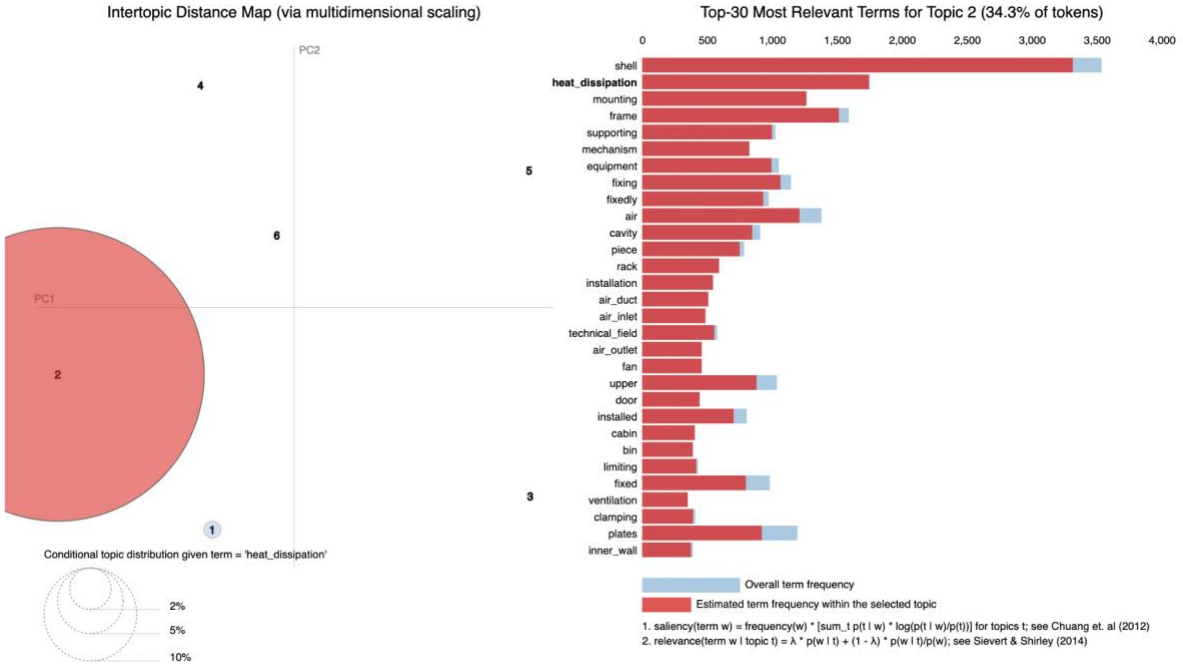
Figure 10: *Intertopic Distance Map and Top-30 Most Relevant Terms for Topic 2.*



Note: *This figure visualizes Topic 2 from the LDA model, which accounts for 34.3% of all tokens in the Y02E60/10 filtered dataset. The intertopic distance map (left) shows the semantic relationships among six topics, while the bar chart (right) highlights Topic 2's top-30 most relevant terms using λ = 0.6. Key terms such as shell, heat_dissipation, and mounting indicate a strong focus on battery enclosures and thermal management systems.*

The next visualization in Figure 11 shows an example of the interactive process of reviewing subject-specific words from the LDAvis output. In this example, the term 'heat_dissipation' has been selected for closer examination. The result shows that the term is mostly related to T2 but can also be related to nearby T1 with a small frequency (indicated by the small blue bubble). The term cannot be linked to any of the other topics where the frequency is zero. The comparison between the red bars (frequency of the term within a specific topic) and the gray bars (total frequency of the term in the entire corpus) can give an idea of the extent to which the terms occur and overlap with each other. For T2, terms such as 'air_duct,' 'air_inlet,' or 'air_outlet' occur only in T2 and are therefore topic-specific.

Figure 11: *Interactive Topic-Specific Term Exploration in LDAvis: Example from Topic 2.*



Note: *This figure exemplifies the interactive review of subject-specific vocabulary using LDAvis, as referenced in the main text. The left panel displays the Intertopic Distance Map, where Topic 2 (T2) emerges as the most dominant and isolated in semantic space. Upon selecting the term heat_dissipation, its conditional distribution is visualized—showing it is predominantly associated with T2, with minor relevance to T1 and none to other topics.*

This approach was replicated for all topics and terms, and the comprehensive visualization via LDAvis is documented in Appendix 2. A summary of the analysis is provided in Table 14, which breaks down each topic's token share (shows the percentage of all words in the dataset that the model assigns to a particular topic), selected terms of interest in the technical domain, and a more concise description of what the topic terms cover. The intertopic distance map revealed that T2 occupied the largest token share, accounting for 34.3% of all tokens (words) in the corpus. This suggests that concepts related to battery enclosures and thermal dissipation were not only frequent but also thematically central across the dataset. In contrast, T6 held a token share of 9.2%, indicating a more specialized and narrowly scoped theme related to safety modules and stack integration.

Table 14: *Topic Summaries Derived from LDAvis.*

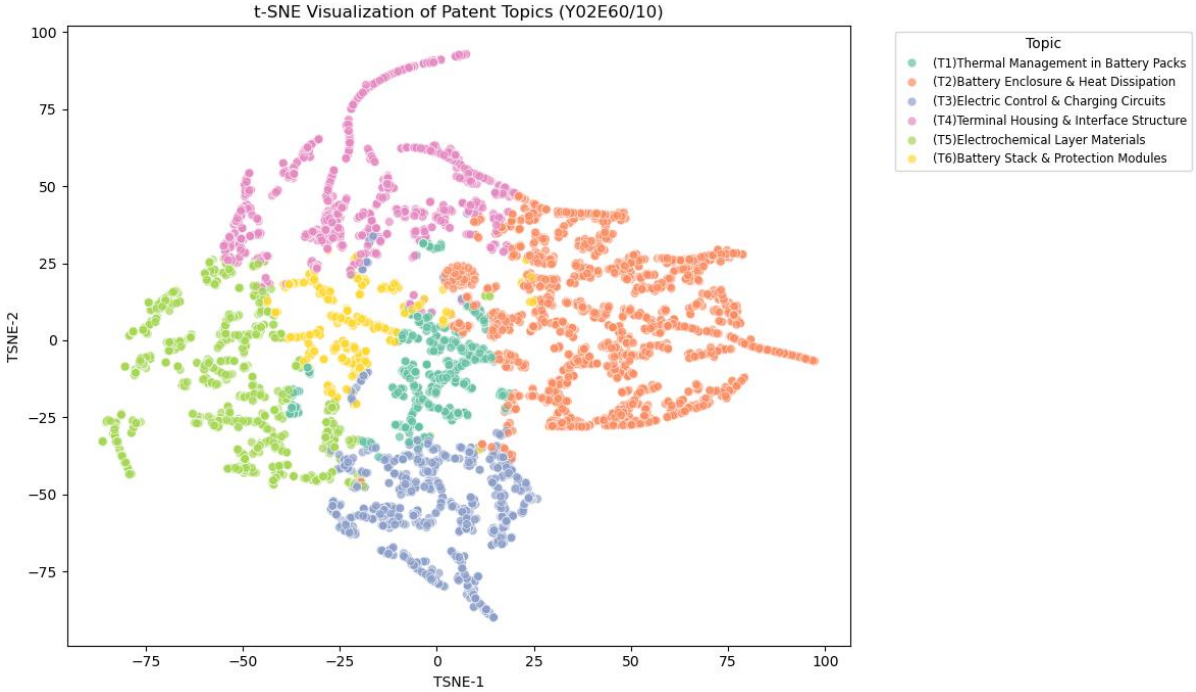
Topic	Token Share	Relevant Keywords (λ=0.6)	Thematic Focus
1	12.4%	cooling, liquid, pipeline, heat, heat_exchange, inlet, outlet, thermal_management, liquid_cooling	Liquid-based cooling systems and thermal regulation
2	34.3%	shell, heat_dissipation, mounting, supporting, air, air_duct, ventilation, rack, equipment	Heat dissipation, air-based cooling, and equipment mountings
3	14.0%	charging, voltage, current, controller, circuit, output, management, monitoring, converter, dcdc_converter	Electric power regulation, charging infrastructure, and control electronics
4	16.1%	portion, terminal, member, direction, bus_bar, current_collecting, adhesive_layer, external_terminal	Terminal architecture, current collection, and electrical connectivity
5	14.0%	electrolyte, active_material, electrochemical, cathode, anode, separator, carbon, electrodes, polymer	Chemical and material components of electrochemical cells
6	9.2%	modularized, activated_carbon, silica_gel, board, stack, motor_vehicle, spring, conductive_carbon, unitary	Modular battery casings with conductive carbon enhancements

Note: *This table summarizes the six dominant topics identified through Latent Dirichlet Allocation (LDA) on the Y02E60/10 green patent corpus. The "Token Share" indicates the percentage of all words (tokens) in the dataset that belong to each topic. "Relevant Keywords" reflect the top topic-specific terms selected using a λ-value of 0.6, which emphasizes terms highly representative of the topic rather than merely frequent across the corpus. The final column provides a concise thematic interpretation, offering a human-readable summary of each topic's technological scope. Full visualizations for each topic are available in Appendix 3.*

5.5 t-SNE Visualization Result

To facilitate the interpretation of the results from the subject modeling, this study employed t-distributed stochastic neighbor embedding (t-SNE) to visualize how individual patent documents are grouped based on their subject distribution (θ matrix) from the LDA model. In contrast to previous visualizations from LDAvis, in which bubbles represented topics and their keywords, the t-SNE method focuses on visualizing individual patents in relation to their dominant topic. In other words, each point in the diagram (Figure 12) represents a patent, and the color of the point corresponds to the patent's dominant topic. The two-dimensional projection facilitates the observation of the semantic similarity between documents, with documents that are proximate to each other exhibiting greater similarity in their topic composition, while those that are distant from each other demonstrate increased distinctness. The t-SNE diagram reveals six visually distinct clusters, with each cluster corresponding to a topic identified by LDA. The distinctiveness of these clusters indicates that the topic model has effectively distinguished between disparate domains of technical innovation in the battery field. For instance, T2 constitutes the most substantial and cohesive group, thereby emphasizing its pivotal function in contemporary patent activity within the battery domain.

Figure 12: *t-SNE Visualization of Patent Topic Distributions Based on LDA Model.*



Note: *This figure presents a two-dimensional t-SNE projection of the topic distributions (θ-matrix) derived from Latent Dirichlet Allocation (LDA) applied to patents classified under CPC Y02E60/10. Each dot represents a single patent, colored according to its dominant topic. The spatial proximity between dots indicates semantic similarity in topic composition, with visible clusters corresponding to distinct areas of innovation such as thermal management, electrochemical materials, or modular protection systems. The separation and density of these clusters demonstrate the coherence and distinctiveness of the LDA-derived topics.*

The distance between clusters in Figure 12 is indicative of conceptual differentiation. For instance, T1 (heat management in battery packs) is more closely aligned with T2, which is logical since both relate to strategies for heat control in batteries, although in different approaches (liquid-based versus structural design). The existence of overlap between clusters suggests the potential for patents to encompass multiple themes. For instance, T6 exhibits a pattern of individual patents that intersect with all subject clusters. The density of points within clusters also provides useful information: areas with higher density may indicate a concentration of patents with very similar content, while areas with lower density may point to more exploratory or new innovations.

The quality of the clustering was assessed by calculating a silhouette value of 0.4736. This measure is employed to compare the mean similarity within clusters with the mean difference between clusters. A value approaching 0.5 signifies an adequate balance between the compactness and separation of the cluster. In practice, this means that most patents are well grouped with others that have similar content, confirming that the subjects are not only statistically related but also interpretable in real terms (Rousseeuw, 1987).

The efficacy of the T-SNE map is further enhanced through meticulous parameter tuning. In accordance with recent best practices (Gove et al., 2022), three configurations were examined to optimize visualization, with variations in perplexity, exaggeration, and learning rate. The final version employed PCA-based initialization and Barnes-Hut approximation to enhance stability and computational efficiency (van der Maaten & Hinton, 2008; scikit-learn, n.d.). Following an examination of the various configurations, option A (see Table 12) was selected, given its superior silhouette score. This configuration was determined to be optimal based on the settings perplexity=30, early exaggeration=12, and learning rate=auto.

In summary, this t-SNE visualization not only confirms the effectiveness of the topic modeling but also facilitates the interpretation of the distribution and thematic structure of battery innovations in a highly visual and accessible format. The LDA model is augmented by this valuable complement, which demonstrates the aggregation of technical concepts and offers specific insights into how individual patents emerge as a complement to the LDA visualization.

6. Discussion

This chapter interprets the results in relation to the research questions and theoretical framework. It discusses the relevance and limitations of the extracted topics, compares them with known market and technological trends, and assesses the utility of LDA and S-curve modeling in capturing innovation maturity. The discussion also addresses the implications for future research and highlights how semantic patterns in patent texts reflect technological trajectories and industrial focus areas.

6.1 Patent Trend Discussion

The analysis of patent trends in this study is based directly on the technology life cycle (TLC) model and the theory of innovation diffusion introduced in Chapter 3 (Fleming, 2001; Tushman & Anderson, 1986; Haupt et al., 2007). In line with these frameworks, the logistic S-curve was chosen as the tool for modeling the temporal development of technologies represented by CPC codes. The choice was theoretically motivated by the TLC assumption that technologies follow a sigmoidal trajectory, with initial experimentation, rapid growth followed by stabilization, which is well aligned with the shape of a logistic function.

By applying the method to patent data, the S-curve in this study could be used to assess the different phases of technologies and thus identify those that are in growth. For example, Y02E60/10 (battery storage systems) showed exponential growth from 2015 onwards, which is consistent with the growth phase described by Tushman and Anderson (1986). The model's inflection point around 2024 indicates a likely transition to maturity, a point at which companies typically consolidate around dominant designs (Abernathy & Utterback, 1978). In contrast, the broader category Y02E60 showed signs of earlier saturation, reflected in its flattening curve, which is characterized by incremental innovation dominating research and development in a later stage of the TLC (Utterback and Abernathy, 1975).

However, it is important to note that the method is not without its limitations. As Haupt et al. (2007) have noted, the number of patents does not always accurately reflect the technical potential or disruption, particularly in areas where incentives for patenting are not uniform. In addition, the S-curve model assumes dependence on previous development and continuity, assumptions that can be broken in pioneering sectors or, for example, in the event of regulatory shocks. Therefore, although the S-curve is consistent with the TLC and supports the study's aim of mapping innovation maturity, it should be interpreted as a guiding rather than a deterministic tool.

6.2 Interpretation of Topic Findings

Batteries are a key component in the global transition to clean energy. Whether used in electric vehicles (EVs), stationary grid storage systems, or consumer electronics, innovations in battery technology are crucial to reducing emissions, increasing energy efficiency, and improving the resilience of energy systems. As demand for batteries increases, so does the need to understand where innovation is taking place and what problems engineers are working to solve. In this thesis, LDA has been used to analyze patents classified under CPC Y02E60/10, a code specific to energy storage with batteries. Six topics emerged from the LDA analysis with the highest coherence scores, each representing a specific area of technological development. These topics are not only technically distinct but also reflect real-world challenges in making batteries safer, more efficient, and more adaptable.

The topics of “heat management” (T1) and “enclosure and heat dissipation” (T2) can be connected to the two primary methods used to ensure a safe temperature in batteries, with the utilization of liquids and air. T1 generated keywords such as "cooling," "liquid,” "heat exchange," and "thermal management." These terms indicate a strong focus on liquid cooling

systems in batteries, which are important for regulating operating temperatures (Pesaran, 2002; Wu et al., 2019). Liquid cooling is widely used in high-performance applications such as EVs and grid storage, where thermal stability is essential to safety and longevity. Roe et al. (2022) confirm that immersion and plate-based liquid cooling have become key research areas, as these systems improve heat removal while supporting fast charging. The steady rise in this topic over time aligns with the growing demand for higher-density battery packs and faster charging protocols. Concurrently, T2 generates terms such as “heat dissipation,” “air,” “ventilation,” and “frame,” focusing on air-based cooling and the physical design of battery enclosures. As posited by Wu et al. (2002), the utilization of enclosures for the purpose of air cooling is a comparatively uncomplicated process. Nevertheless, these enclosures have been observed to encounter difficulties in operating within high-power conditions. Zhang et al. (2023) and Han et al. (2025) highlight recent innovations in enclosure geometry and airflow optimization that enhance cooling without added complexity. Air-cooled battery modules remain prevalent in less complex electric cars and stationary systems, where cost-effectiveness and moderate heat loads prevail (Wu et al., 2019). The recent spike in patents under this topic suggest a resurgence in mechanical design as a crucial tool in improving battery safety and performance, especially in low-to-mid-power applications.

“Charging circuits” (T3) and “terminal interfaces” (T4) can be assimilated with the electrical control and connection of batteries. T3 incorporated keywords such as "charging," "voltage," "control unit," "current," and "management.” This topic reflects innovations in charging control and battery management systems (BMS). As EV infrastructure expands and vehicle-to-grid (V2G) systems gain traction, intelligent charge regulation becomes increasingly important. Rana et al. (2024) note that the rise of V2G charging requires more robust and responsive BMS architectures to manage power flows safely and efficiently. These systems are critical for regulating energy flow, balancing cells, and ensuring safe operation (Azooz & Sulayman, 2007). The growth in this topic indicates that advanced electronics and control systems remain at the forefront of battery technology. Furthermore, T4 generated keywords such as "terminal," "surface," "housing," "busbar," and "current collector" that focus on physical connections and electrical interfaces within battery modules. Poor contact quality has been demonstrated to result in the formation of hotspots, energy loss, and accelerated degradation (Saxon et al., 2024). As higher-current applications become more common, connection quality has emerged as a key determinant of safety and efficiency, validating this topic’s relevance.

“Electrochemical materials” (T5) and “stack and protection modules” (T6) focus on the core of the battery, its chemistry and design. T5 incorporated keywords such as “electrolyte,” “active material,” “cathode,” “anode,” and “separator,” emphasizing pivotal battery materials and chemistry. Material advancements are foundational to increasing energy density, safety, and charging speed. Kim et al. (2020) emphasize the development of high-nickel cathodes, silicon anodes, and solid electrolytes as active research areas. Advancements in solid electrolytes, high-nickel cathodes, and silicon anodes have led to substantial enhancements in energy density and safety (Tarascon & Armand, 2001; Goodenough & Park, 2013). Moreover, T6 concentrated on “stack,” “housing,” “storage,” and “modulation,” with an emphasis on modular battery designs and thermal protection systems. Liu et al. (2014) demonstrated that modular stacks present significant challenges in achieving uniform thermal management. Modularization has been demonstrated to facilitate maintenance and recycling; however, if not designed with sufficient care, it can pose a risk of electrochemical and thermal imbalances (Liu et al., 2014). The growing prevalence of this topic signals a shift toward flexible, serviceable, and safer system-level battery solutions.

6.3 Reflection on Model Effectiveness

This study examined green patents within the CPC classification Y02E60/10, with a particular focus on energy storage utilizing batteries. The application of LDA to a corpus of patent documents resulted in the identification of six topics. The study adopts an exploratory approach, in which the outcome and choice of technological domain were unclear beforehand. Consequently, no predetermined theoretical frameworks were available for the purpose of comparing the results. Instead, the purpose of the discussion section was to highlight the findings in the results and to contextualize them. During the course of the project, it proved challenging to evaluate the relevance of the results obtained from the various topics, primarily due to a lack of experience and expertise in the field of these technologies. It was evident that this research method holds value when the researcher possesses familiarity with the technological aspects under investigation and when working in close collaboration with domain experts who can evaluate the relevance of the topics (Chang et al., 2009).

The LDA model, supported by LDAvis and t-SNE visualizations, successfully isolated conceptually distinct themes, which are both semantically and technologically meaningful. Each topic addresses a real challenge in battery development, from thermal regulation and material chemistry to mechanical design and electrical integration. The minimal overlap

between topics in the intertopic distance map suggests strong separation and coherence, lending credibility to the modeling approach.

The decision to utilize LDA for thematic analysis was made based on its capacity to identify latent semantic structures from unstructured text (Blei et al., 2003). However, as demonstrated in the existing literature, several alternative or complementary methods have been employed in patent analysis, each with its own advantages and disadvantages.

In their 2021 study, Arts et al. propose a more direct, text-based method based on natural language processing (NLP). This method involves tracking keywords, bigrams, trigrams, and new concept combinations from patent titles, abstracts, and claims over time to analyze how technical elements are reused and disseminated. This approach diverges from LDA in that it captures actual semantic reuse rather than probabilistic latent themes, thereby offering potentially higher resolution for identifying groundbreaking innovations. As this study did not originate from a predetermined technical domain or a predefined set of keywords, the absence of predetermined terminology promoted the unsupervised nature of LDA. LDA enabled the inductive detection of dominant topics, which aligned with the exploratory, data-driven research design underlying this study (Bryman & Bell, 2019).

In terms of forecasting, the logistic S-curve remains a theoretically grounded model that is consistent with the technology life cycle framework (Haupt et al., 2007; Tushman & Anderson, 1986). The S-curve has proven to be a valuable tool for identifying saturation points and modeling the cumulative progression of technological development. However, it is important to note that the S-curve operates under the assumption of a steady, monotonic spread. This may limit its applicability in sectors that are prone to disruptive innovations or regulatory measures.

In summary, alternative methods offer valuable perspectives, especially in more focused or well-defined areas. However, the combined use of LDA and logistic S-curves in this thesis provided an appropriate balance between exploratory breadth, theoretical fit, and relevance for management. Subsequent research can build on this foundation by integrating hybrid models or triangulating LDA results with semantic tracking techniques such as those proposed by Arts et al. (2021).

7. Conclusion

This thesis set out to explore the thematic structures and innovation maturity within the domain of climate change mitigation technologies, using large-scale patent data as a foundation. By integrating Latent Dirichlet Allocation (LDA) topic modeling with logistic S-curve analysis, the study contributes a novel methodological approach for mapping and interpreting patterns of green technological development.

Initially, an examination of patent applications within adoption and climate change mitigation technologies in CPC classification Y02 revealed that the most prevalent group was for renewable technologies (Y02E10). The findings, as indicated by the logistic S-curve, revealed that all subgroups had plateaued in their growth trajectory. Stagnant growth may be defined as an era of diminished innovation, characterized by a tendency towards standardization and the emergence of a dominant design paradigm. This period is often marked by a shift towards cost-effective innovation as opposed to the development of revolutionary ideas. The second most frequent CPC code, the subgroup for battery-based energy storage technologies (Y02E60/10), was found to still be in the growth phase. In the hope of finding greater innovation variation, this subgroup was selected for further analysis using subject modeling.

The central research question—What are the dominant thematic structures in patents for climate change mitigation? —was addressed through an unsupervised machine learning analysis of 7,211 patents classified under the CPC categories Y02E60/10. The LDA modeling revealed six coherent topics (coherence = 0.6945) that reflect different technical areas, including battery cooling systems, chemical composition, safety modulation, and design. The trend demonstrated that topic 2, which encompasses technologies for managing battery temperature with air, has undergone exponential growth since 2020 and is currently the most trending topic.

The initial sub-question pertained to the efficacy of employing unsupervised topic modeling in identifying innovation trends. The findings indicate that LDA, when implemented on preprocessed patent abstracts, is effective in unveiling latent structures that correspond to real-world technical domains.

The second sub-question centered on interpretability and validation. The utilization of visualization techniques, including LDAvis and t-SNE, proved to be instrumental in clarifying the relationships between topics and patents. t-SNE revealed clear clusters of topics with a silhouette score of 0.4736, while LDAvis facilitated the review of prominent and distinct terms.

The efficacy of these methodologies was demonstrated by their significant contribution to enhancing the transparency and topic separation of the model.

From a theoretical perspective, the study extends the application of topic modeling and S-curve frameworks in the analysis of innovation maturity. In practice, the results provide useful insights for decision-makers, technology managers, and R&D strategists. Patent documents, which are often underutilized due to their complexity, were found to contain rich and analyzable content that can inform decisions regarding technology investments, competitive positioning, and policy priorities. The findings underscore the growing importance of data-driven forecasting tools in innovation planning.

In summary, this thesis demonstrates the value of combining text mining and innovation theory to extract meaningful insights from patent corpora. As green innovation continues to accelerate in response to global climate goals, such methods will be crucial for navigating the complexity of technological change and supporting evidence-based innovation strategies.

7.1 Limitations and Future Research

This study presents a comprehensive analysis of climate mitigation patents using LDA topic modeling and logistic S-curve forecasting. However, it has important limitations, many of which point directly to meaningful avenues for future research.

First, the dataset was limited to 50,000 patent records due to technical export restrictions, although efforts were made to ensure temporal representativeness; this is considered a limitation on generalizability. Future research could include a larger sample to provide better representation of the population. Furthermore, the analysis does not consider jurisdictional differences or patent family relationships. This limits the ability to assess geographical differences in innovation dynamics, a significant shortcoming given the role of national policy in climate innovation (e.g., subsidies or emission targets). Future studies should explore regional samples or jurisdiction comparisons to address this limitation.

Second, although LDA successfully revealed interpretable thematic clusters, the labeling process from keywords to AI-generated labels was based on expert validation rather than validation by experts in the field. This trade-off, driven by time and resource constraints, may have introduced interpretative bias. Future work should include expert panels or semi-supervised methods (e.g., human-assisted topic validation) to improve the validity and technical nuances of the labeling.

Third, although based on life cycle theory, the application of the S-curve model assumes a continuous and cumulative innovation trajectory. However, this assumption may not hold in sectors prone to sudden regulatory changes or technological disruptions. As discussed in Haupt et al. (2007), the logistic model is particularly challenged by overlapping generations of technological development. Future research should experiment with ensemble trend models (e.g., a combination of S-curve and Prophet) or leverage technology substitution models to account for such dynamics.

Finally, this study focused on a single CPC subclass (Y02E60/10) for thematic insights. While this provided clarity, it limits the results to the broader Y02 area. Future work could apply comparisons between subclasses or investigate patterns of technology convergence across areas to discover systemic innovation pathways relevant to climate goals.

References

Adams, J. N. (2019). *History of the patent system*. In Research handbook on patent law and theory (pp. 2–26). Edward Elgar Publishing.

Allison, J. R., Lemley, M. A., Moore, K. A., & Trunkey, R. D. (2003). Valuable patents. *Georgetown Law Journal, 92*, 435.

Altuntas, S., Dereli, T., & Kusiak, A. (2015). Forecasting technology success based on patent data. *Technological Forecasting and Social Change, 96*, 202–214.

Arts, S., Appio, F. P., & Van Looy, B. (2013). Inventions shaping technological trajectories: Do existing patent indicators provide a comprehensive picture? *Scientometrics, 97*(3), 397–419. https://doi.org/10.1007/s11192-013-1045-1

Arts, S., Hou, J., & Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. Research policy, 50(2), 104144.

Azooz, A. A., & Sulayman, A. H. (2007). A simple method for improving energy efficiency in battery charging. *Journal of Applied Sciences Research, 3*(12), 1543–1547.

Benson, C. L., & Magee, C. L. (2015). Quantitative determination of technological improvement from patent data. *PLOS ONE, 10*(4), e0121635. https://doi.org/10.1371/journal.pone.0121635

Bergeaud, A., Potiron, Y., & Raimbault, J. (2017). Classifying patents based on their semantic content. *PLOS ONE, 12*(4), e0176310.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Cai, W., & Xu, F. (2021). The impact of the new environmental protection law on eco-innovation: Evidence from green patent data of Chinese listed companies. *Environmental Science and Pollution Research, 29*, 10047–10062. https://doi.org/10.1007/s11356-021-16365-1

Chang, Y., & Cheng, Q. (2024). Entrepreneurial mentoring, financial support, and incubator patent licensing: Evidence from Chinese incubators. *European Journal of Innovation Management, 27*(1), 290–309. https://doi.org/10.1108/EJIM-03-2022-0140

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 22, 288–296.

Chuang, J., Manning, C. D., & Heer, J. (2012, May). Termite: Visualization techniques for assessing textual topic models. In Proceedings of the international working conference on advanced visual interfaces (pp. 74-77).

de Lange, D. E. (2024). Climate action now: Energy industry restructuring to accelerate the renewable energy transition. *Journal of Cleaner Production, 443*, 141018. https://doi.org/10.1016/j.jclepro.2024.141018

Desheng, L., Jiakui, C., & Ning, Z. (2021). Political connections and green technology innovations under an environmental regulation. *Journal of Cleaner Production.* https://doi.org/10.1016/

Du, K., Li, P., & Yan, Z. (2019). Do green technology innovations contribute to carbon dioxide emission reduction? Empirical evidence from patent data. *Technological Forecasting and Social Change, 146*, 297–303. https://doi.org/10.1016/j.techfore.2019.06.010

Energy Institute. (2024). *Statistical Review of World Energy 2024* (73rd edition). London: Energy Institute. https://www.energyinst.org/statistical-review

European Patent Office (EPO) & United States Patent and Trademark Office (USPTO). (2013). *Cooperative Patent Classification (CPC) scheme and definitions*. https://www.cooperativepatentclassification.org

EPO & USPTO. (n.d.). About CPC. *Cooperative Patent Classification*. https://www.cooperativepatentclassification.org/about

Favot, M., Vesnic, L., Priore, R., Bincoletto, A., & Morea, F. (2023). Green patents and green codes: How different methodologies lead to different results. *Resources, Conservation & Recycling Advances, 18*, 200132. https://doi.org/10.1016/j.rcradv.2023.200132

Fischer, T., & Leidinger, J. (2014). Testing patent value indicators on directly observed patent value—An empirical analysis of Ocean Tomo patent auctions. *Research Policy, 43*(3), 519–529.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science, 47*(1), 117–132.

Fleming, L., Mingo, S., & Chen, D. (2007). Collaborative brokerage, generative creativity and creative success. *Administrative Science Quarterly, 52*(3), 443–475.

Gambardella, A., Harhoff, D., & Verspagen, B. (2008). The value of European patents. *European Management Review, 5*(2), 69–84.

Gan, J., & Qi, Y. (2021). Selection of the optimal number of topics for LDA topic model—taking patent policy analysis as an example. Entropy, 23(10), 1301.

Gao, L., Porter, A., Wang, J., Fang, S., Zhang, X., Ma, T., Wang, W., & Huang, L. (2013). Technology life cycle analysis method based on patent documents. *Technological Forecasting and Social Change, 80*, 398–407. https://doi.org/10.1016/j.techfore.2012.10.003

Goodenough, J. B., & Park, K. S. (2013). The Li-ion rechargeable battery: A perspective. Journal of the American Chemical Society, 135(4), 1167–1176. https://doi.org/10.1021/ja3091438

Gove, R., Cadalzo, L., Leiby, N., Singer, J. M., & Zaitzeff, A. (2022). New guidance for using t-SNE: Alternative defaults, hyperparameter selection automation, and comparative evaluation. *Visual Informatics, 6*(1), 87–97. https://doi.org/10.1016/j.visinf.2022.04.003

Hall, B. H. (2009). Business and financial method patents, innovation, and policy. *Scottish Journal of Political Economy, 56*(4), 443–473. https://doi.org/10.1111/j.1467-9485.2009.00493.x

Hall, B. H., & Harhoff, D. (2012). Recent research on the economics of patents. *Annual Review of Economics, 4*(1), 541–565.

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2000). Market value and patent citations: A first look. *NBER Working Paper No. 7741.*

Han, Y., Bai, Y., Zhang, W., Cui, S., Zang, L., & Ding, H. (2025). Study of thermal management system for battery box for formula student electric racing car. *Frontiers in Mechanical Engineering, 11*, 1529633. https://doi.org/10.3389/fmech.2025.1529633

Haupt, R., Kloyer, M., & Lange, M. (2007). Patent indicators for the technology life cycle development. *Research Policy, 36*(3), 387–398.

Hughes, J. (1988). The philosophy of intellectual property. *Georgetown Law Journal, 77*, 287.

International Renewable Energy Agency (IRENA). (2024). *Renewable capacity statistics 2024*. Abu Dhabi: IRENA. https://www.irena.org/Publications

Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal, 36*(10), 1435–1457.

Kim, Y. K., Lee, K., Park, W. G., & Choo, K. (2012). Appropriate intellectual property protection and economic growth in countries at different levels of development. *Research Policy, 41*(2), 358–375.

Kim, H.-J., Krishna, T. N. V., Zeb, K., Rajangam, V., Gopi, C. V. V. M., Sambasivam, S., & Raghavendra, K. V. G. (2020). A Comprehensive Review of Li-Ion Battery Materials and Their Recycling Techniques. *Electronics, 9*(7), 1161. https://doi.org/10.3390/electronics9071161

Klepper, S. (1996). Entry, exit, and innovation over the product life-cycle. *American Economic Review, 86*(3), 562–583.

Kürtössy, J. (2004). Innovation indicators derived from patent data. *Periodica Polytechnica Social and Management Sciences, 12*(1), 91–101.

Lens. (2025a). *Collective action project. Lens [CAP]*. https://www.lens.org/lens/collective-action

Lens. (2025b, February 11). *What is the Lens*. https://about.lens.org/what/

Lemley, M. A., & Feldman, R. (2016). Patent licensing, technology transfer, and innovation. *American Economic Review, 106*(5), 188–192.

Li, Q., Maggitti, P. G., Smith, K. G., Tesluk, P. E., & Katila, R. (2013). Top management attention to innovation: The role of search selection and intensity in new product introductions. *Academy of Management Journal, 56*(3), 893–916.

Li, S., Hu, J., Cui, Y., & Hu, J. (2018). DeepPatent: Patent classification with convolutional neural networks and word embedding. *Scientometrics, 117*(2), 721–744.

Li, Z., Tate, D., Lane, C., & Adams, C. (2012). A framework for automatic TRIZ level of invention estimation of patents using natural language processing, knowledge-transfer and patent citation metrics. *Computer-Aided Design, 44*(10), 987–1010.

Liu, R., Chen, J., Xun, J., Jiao, K., & Du, Q. (2014). Numerical investigation of thermal behaviors in lithium-ion battery stack discharge. Applied Energy, 132, 91–102. https://doi.org/10.1016/j.apenergy.2014.06.042

Mansfield, E. (1986). Patents and innovation: An empirical study. *Management Science, 32*(2), 173–181.

Marín-Vinuesa, L. M., Scarpellini, S., Portillo-Tarragona, P., & Moneva, J. M. (2020). The impact of eco-innovation on performance through the measurement of financial resources and green patents. *Organization & Environment, 33*(2), 285–310. https://doi.org/10.1177/1086026618819103

Mueller, D. C., & Tilton, J. (1969). Research and development costs as a barrier to entry. *Canadian Journal of Economics, 2*(4), 570–579.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL* (pp. 100–108).

Roe, C., Feng, X., White, G., & Li, R. (2022). Immersion cooling for lithium-ion batteries – A review. *Journal of Power Sources, 525*, 231094. https://doi.org/10.1016/j.jpowsour.2022.231094

O'callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications, 42(13), 5645-5657.

Omar, M., On, B. W., Lee, I., & Choi, G. S. (2015). LDA topics: Representation and evaluation. *Journal of Information Science, 41*(5), 662–675. https://doi.org/10.1177/0165551515587839

Pan, X., & Xu, Y. (2023). Advancements of artificial intelligence techniques in the realm about library and information subject—A case survey of Latent Dirichlet Allocation method. *IEEE Access, 11*, 132627–132634. https://doi.org/10.1109/ACCESS.2023.3334619

Park, H., Yoon, J., & Kim, K. (2013). Identification and evaluation of corporations for merger and acquisition strategies using patent information and text mining. *Scientometrics, 97*, 883–909.

Park, I., & Yoon, B. (2014). A semantic analysis approach for identifying patent infringement based on a product–patent map. Technology Analysis & Strategic Management, 26(8), 855-874.

Park, Y., & Yoon, J. (2017). Application technology opportunity discovery from technology portfolios: Use of patent classification and collaborative filtering. *Technological Forecasting and Social Change, 118*, 170–183.

Rainville, A., Dikker, I., & Buggenhagen, M. (2025). Tracking innovation via green patent classification systems: Are we truly capturing circular economy progress? *Journal of Cleaner Production, 413*, 138344. https://doi.org/10.1016/j.jclepro.2024.138344

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399–408). https://doi.org/10.1145/2684822.2685324

Saxon, A., Yang, C., Santhanagopalan, S., Keyser, M., & Colclasure, A. (2024). *Li-Ion Battery Thermal Characterization for Thermal Management Design*. *Batteries*, *10*(4), 136. https://doi.org/10.3390/batteries10040136

Scarpellini, S., Portillo-Tarragona, P., & Marín-Vinuesa, L. M. (2019). Green patents: A way to guide the eco-innovation success process? *Academia Revista Latinoamericana de Administración, 32*(2), 225–243. https://doi.org/10.1108/ARLA-07-2017-0233

scikit-learn. (n.d.). *Tsne*. sklearn.manifold. scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

Serener, B., Kirikkaleli, D., & Addai, K. (2023). Patents on environmental technologies, financial development, and environmental degradation in Sweden: Evidence from novel Fourier-based approaches. *Sustainability, 15*, 302. https://doi.org/10.3390/su15010302

Sichelman, T. (2009). Commercializing patents. *Stanford Law Review, 62*, 341.

Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 952–961).

Sossa, J. W. Z., Marro, F. P., Alzate, B. A., Salazar, F. M. V., & Patiño, A. F. A. (2016). S-Curve analysis and technology life cycle. Application in series of data of articles and patents. Revista ESPACIOS| Vol. 37 (Nº 07) Año 2016.

Tarascon, J. M., & Armand, M. (2001). Issues and challenges facing rechargeable lithium batteries. Nature, 414(6861), 359–367. https://doi.org/10.1038/35104644

Trajtenberg, M., Jaffe, A., & Henderson, R. (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and New Technology, 5*(1), 19–50.

Trappey, A. J., Trappey, C. V., Wu, C. Y., & Lin, C. W. (2012). A patent quality analysis for innovative technology and product development. *Advanced Engineering Informatics, 26*(1), 26–34.

Tushman, M. L., & Anderson, P. (1986). Technological discontinuities and organizational environments. *Administrative Science Quarterly, 31*(3), 439–465. https://doi.org/10.2307/2392832

Pesaran, A. (2002). Battery thermal models for hybrid vehicle simulations. *Journal of Power Sources, 110*(2), 377–382. https://doi.org/10.1016/S0378-7753(02)00186-8

United Nations Framework Convention on Climate Change (UNFCCC). (n.d.a). *The Kyoto Protocol*. https://unfccc.int/kyoto_protocol

United Nations Framework Convention on Climate Change (UNFCCC). (n.d.b). *The Paris Agreement*. https://unfccc.int/process-and-meetings/the-paris-agreement

Urbaniec, M., Tomala, J., & Martinez, S. (2021). Measurements and trends in technological eco-innovation: Evidence from environment-related patents. *Resources, 10*, 68. https://doi.org/10.3390/resources10070068

Utterback, J. M., & Abernathy, W. J. (1975). A dynamic model of process and product innovation. *Omega, 3*(6), 639–656. https://doi.org/10.1016/0305-0483(75)90068-7

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.

Wang, W., Li, Y., Lu, N., Wang, D., Jiang, H., & Zhang, C. (2020). Does increasing carbon emissions lead to accelerated eco-innovation? Empirical evidence from China. *Journal of Cleaner Production, 251*, 119690.

Wang, B., Liu, S., Ding, K., Liu, Z., & Xu, J. (2014). Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology. Scientometrics, 101, 685-704.

World Intellectual Property Organization (WIPO). (1971). *International Patent Classification (IPC): Guide and notes*. https://www.wipo.int/classifications/ipc/en/

World Intellectual Property Organization (WIPO). (2024). *Patents highlights. World Intellectual Property Indicators 2024: Highlights*. https://www.wipo.int/web-publications/world-intellectual-property-indicators-2024-highlights/en/patents-highlights.html

World Intellectual Property Organization (WIPO). (2025). *What is intellectual property (IP)?* https://www.wipo.int/about-ip/en/

Wu, W., Wu, W., Yang, Y., & Lin, Z. (2002). Modeling and analysis of thermal behavior of battery packs for electric vehicles. *Journal of Power Sources, 104*(1), 206–214. https://doi.org/10.1016/S0378-7753(01)00935-1

Wu, W., Xu, J., Liu, Y., & Zhang, X. (2019). Thermal management systems for batteries: A review. *Energy Conversion and Management, 182*, 262–281. https://doi.org/10.1016/j.enconman.2018.12.058

Zhang, C., Wang, Q., Li, K., & Ling, Z. (2023). A Review of Cooling Technologies in Lithium-Ion Power Battery Thermal Management Systems for New Energy Vehicles. *Processes, 11*(12), 3450. https://doi.org/10.3390/pr11123450

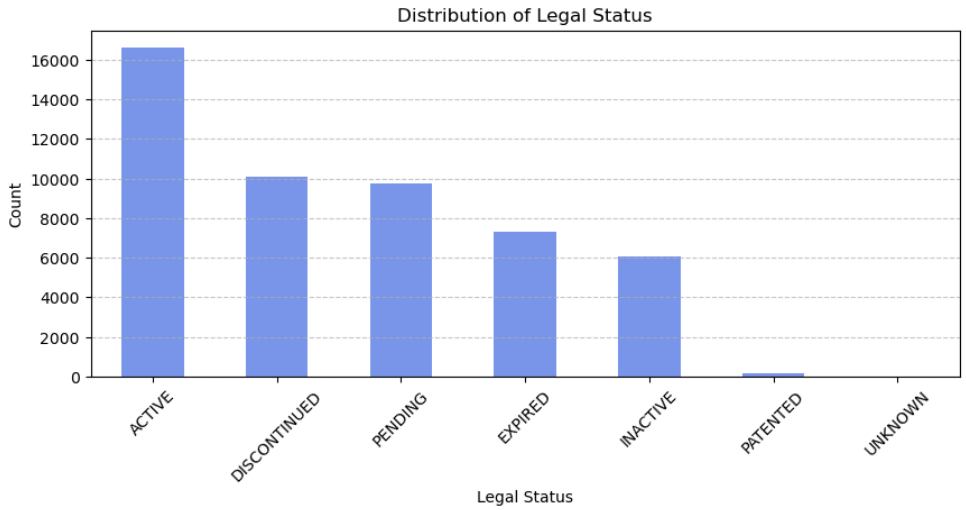
Appendix

This appendix presents a series of visualizations and descriptive analyses that complement the methodological framework outlined in the main chapters of this thesis. The purpose is to transparently document how relevant topics and trends were identified from the green patent dataset, thereby offering deeper insights into the technological focus and evolution of energy storage innovations over time. The analytical process began with the collection of 50,000 patent documents classified under the Y02E60/10 CPC code (energy storage using batteries), retrieved via the Lens.org database. To improve consistency and focus on contemporary innovation patterns, patents published prior to 1980 were removed, resulting in a refined dataset of 48,803 documents. A total of 1,197 entries were filtered out in this step—including 127 forward-dated records listed for 2025. Following data preprocessing (including tokenization, lemmatization, stopword removal, and filtering of non-technical terms), a Latent Dirichlet Allocation (LDA) model was trained to extract interpretable topics from the abstract texts. Each patent was then represented as a probabilistic distribution over the six resulting topics. To validate topic quality and enhance interpretability, the study employed a combination of LDAvis for term relevance and t-distributed stochastic neighbor embedding (t-SNE) to spatially visualize topic separability.

Figures in this appendix illustrate the progression from raw data preprocessing through classification structures, global patent trends, legal status distributions, and temporal trajectories. These are followed by detailed topic modeling outputs including word clouds, citation patterns, top inventors/applicants, and finally pyLDAvis visualizations that explore semantic content and overlap between topics.

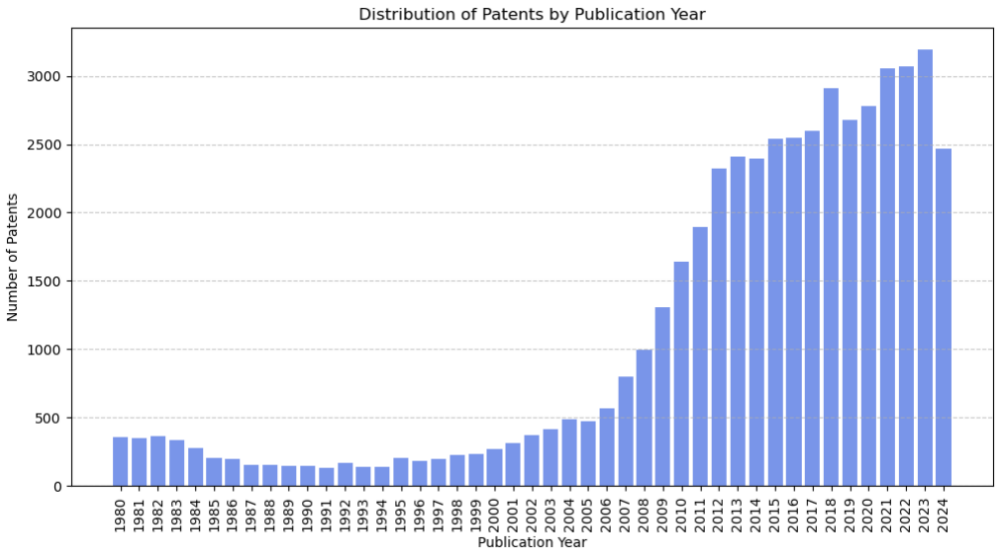
Appendix 1. Overall Dataset Analysis.

Figure 13: *Legal Status Distribution of Patents in the Sample Dataset (n = 50,000).*



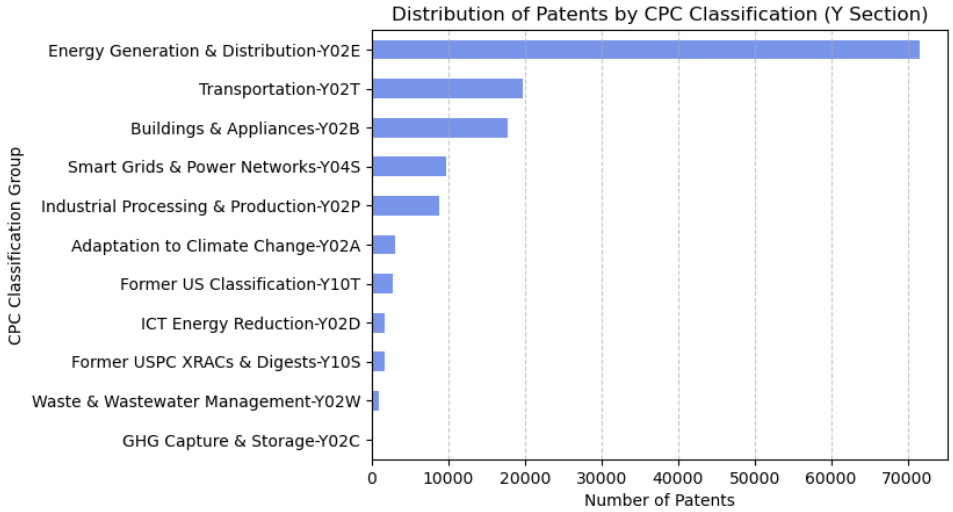
Note: *This bar chart displays the distribution of legal statuses among 50,000 patent records sampled from the Lens.org database. The majority of patents are currently classified as* ***active*** *(n ≈ 16,641), followed by* ***discontinued***, ***pending***, *and* ***expired*** *statuses. Smaller counts are observed for* ***inactive***, ***patented***, *and* ***unknown*** *categories. The data provide insight into the lifecycle stages of green technology patents, indicating that a substantial proportion of innovations are still in force or under examination—highlighting ongoing technological development and potential for future commercialization. This overview supports later assessments of innovation maturity and trend relevance within the corpus.*

Figure 14: *Annual Distribution of Green Patents by Publication Year in the Sample Dataset (1980–2024).*



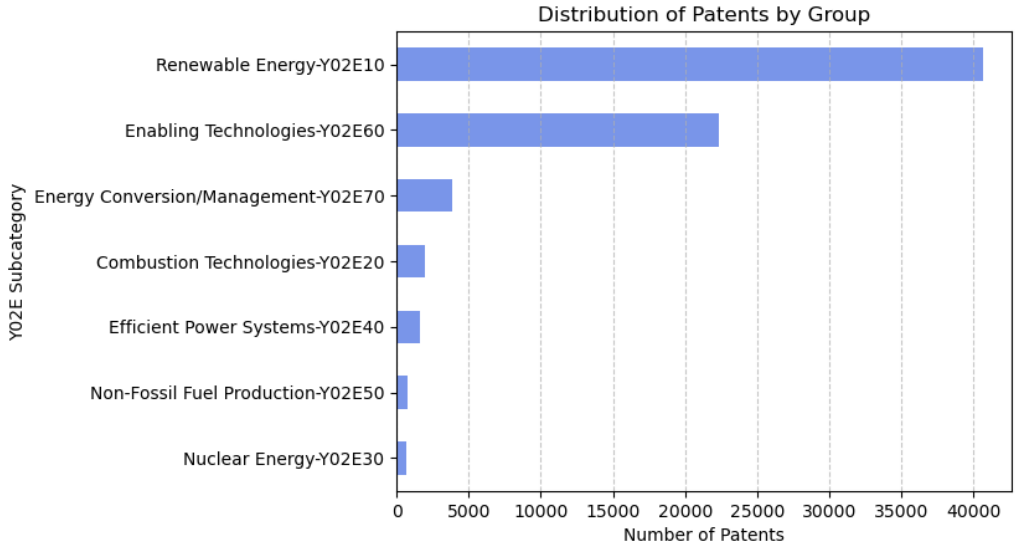
Note: *This bar chart depicts the annual distribution of patent publications (n = 50,000) from 1980 to 2024, based on the subset of green patents retrieved from the Lens.org database. The figure highlights a substantial and steady increase in patenting activity beginning in the early 2000s, with a particularly sharp rise between 2010 and 2023. This upward trend reflects accelerating innovation in energy storage and green technologies, driven by regulatory pressure, market demand, and global decarbonization goals. The drop observed for 2024 is likely due to incomplete data for the most recent year at the time of data collection.*

Figure 15: *Distribution of Patents by CPC Green Technology Classifications (Y Section).*



Note: *This horizontal bar chart displays the distribution of patents across various Cooperative Patent Classification (CPC) groups within the Y section, which covers climate change mitigation technologies. The data clearly shows that the majority of green patents fall under Y02E – Energy Generation & Distribution, followed by Y02T (Transportation) and Y02B (Buildings & Appliances). This reflects the strong innovation focus on renewable energy systems, electric mobility, and energy efficiency in the built environment. The lesser-represented classes—such as GHG capture and ICT energy reduction—indicate emerging or more specialized niches in the green patent landscape. This classification breakdown provides insight into sectoral innovation priorities and thematic concentrations in environmental technologies.*

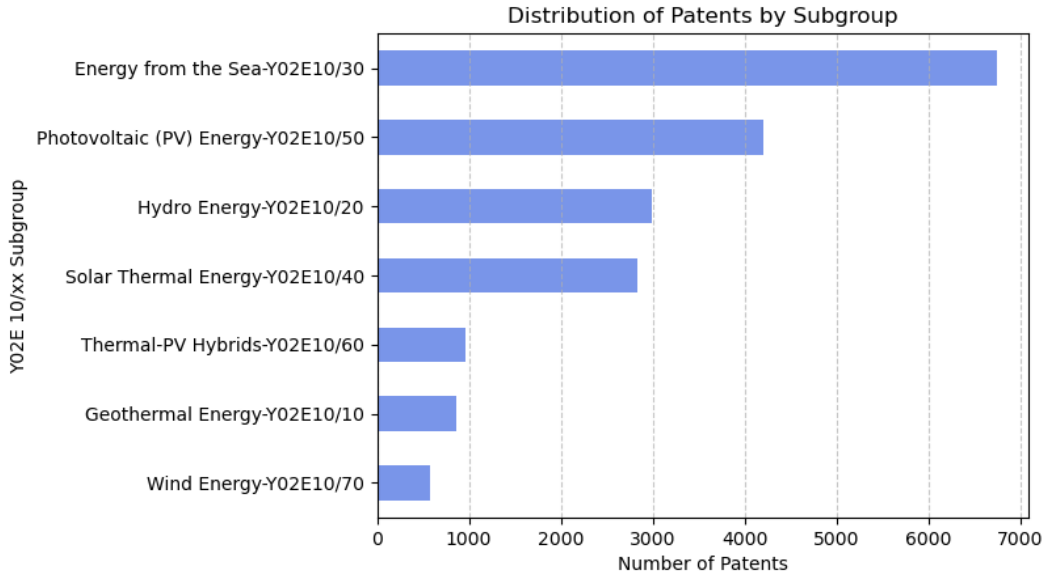
Figure 16: *Distribution of Patents by Y02E Subcategory (Green Technologies—Energy Sector).*



Note: *This bar chart illustrates the distribution of patent filings across subcategories within the Y02E section of the Cooperative Patent Classification (CPC), which focuses on technologies for the generation, transmission, or distribution of energy. The dominant category is Y02E10 – Renewable Energy, followed by Y02E60 – Enabling Technologies, which includes energy storage systems and supporting components. Smaller shares are observed in subgroups such as energy conversion and management (Y02E70), combustion technologies (Y02E20), and nuclear energy (Y02E30). This distribution highlights the industry's prioritization of*

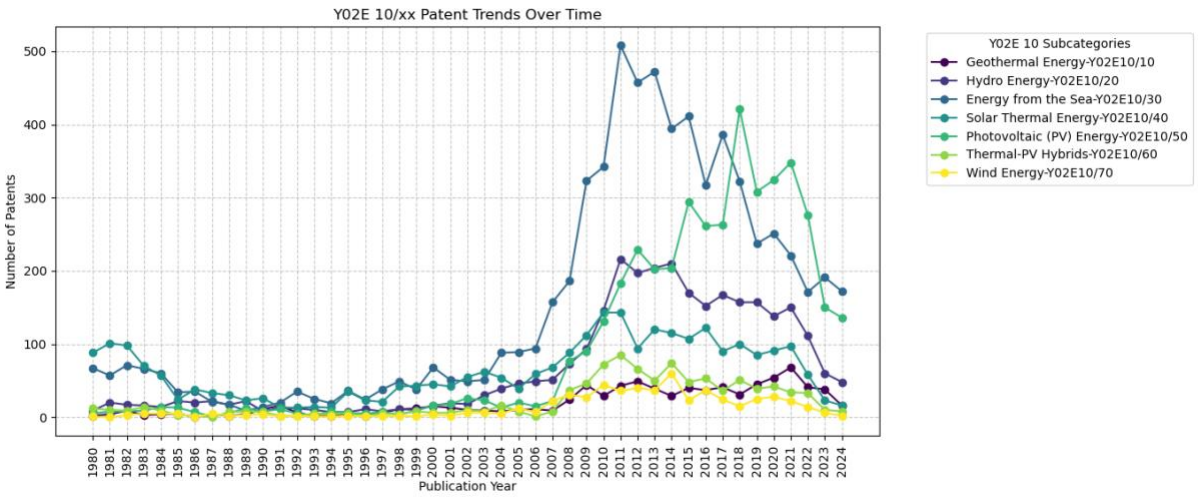
sustainable power generation and infrastructure over more conventional or transitional technologies.

Figure 17: *Distribution of Patents by Renewable Energy Subgroup (Y02E10/xx CPC Classification).*



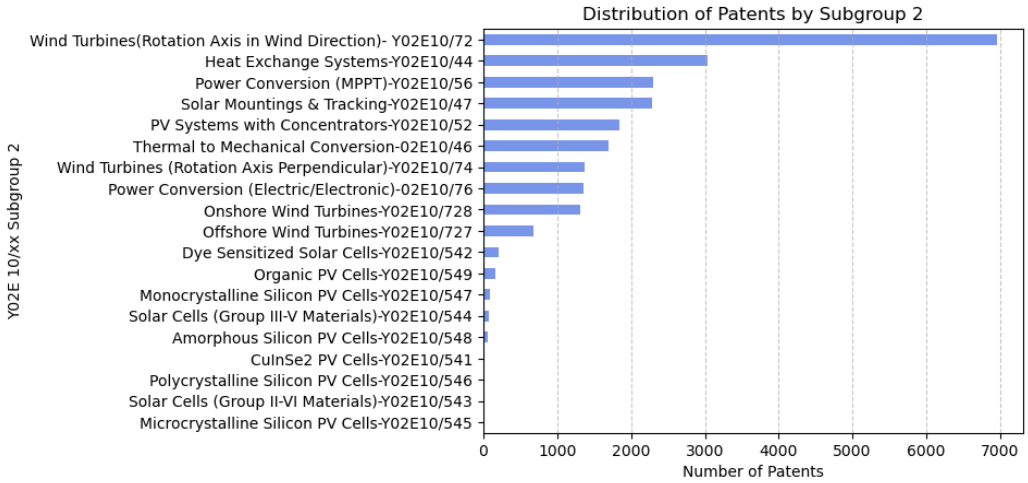
Note: *This figure displays the distribution of patents across subcategories within Y02E10, which pertains to renewable energy technologies in the Cooperative Patent Classification (CPC) system. The two leading subgroups are Energy from the Sea (Y02E10/30) and Photovoltaic (PV) Energy (Y02E10/50), highlighting a strong innovation emphasis on ocean-based and solar power systems. Hydro energy (Y02E10/20) and solar thermal energy (Y02E10/40) also represent significant areas of patent activity. Subgroups such as wind (Y02E10/70), geothermal (Y02E10/10), and hybrid technologies (Y02E10/60) exhibit smaller volumes, suggesting either technological maturity or niche innovation focus. This classification-based breakdown reveals prevailing trends and priorities in renewable energy innovation.*

Figure 18: *Temporal Trends in Renewable Energy Patents by Y02E10/xx Subcategory (1980–2024).*



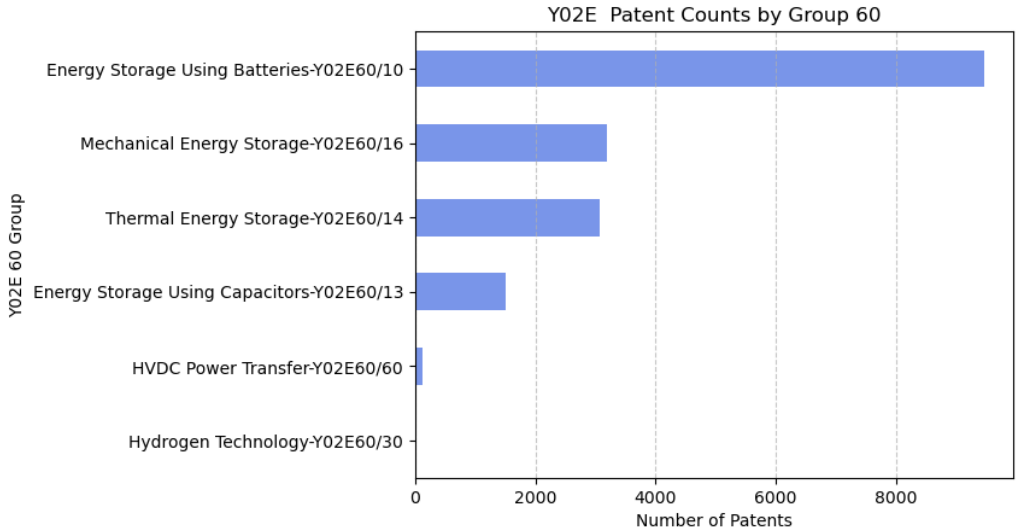
Note: *This line graph illustrates the publication trends over time for patents in renewable energy subcategories under Y02E10 in the CPC classification system. Energy from the Sea (Y02E10/30) and Photovoltaic Energy (Y02E10/50) show the most significant peaks in the early 2010s, reflecting strong innovation interest during that period. While some categories such as Hydro (Y02E10/20) and Solar Thermal (Y02E10/40) maintain moderate consistency, others like Wind (Y02E10/70) and Geothermal (Y02E10/10) remain niche areas with fewer filings. The decline post-2018 may be influenced by both market saturation and patent publication lags. Overall, the figure highlights shifting innovation dynamics in renewable energy technologies over the past four decades.*

Figure 19: *Distribution of Patents by Detailed Subgroup within Y02E10 (Energy Technologies)*



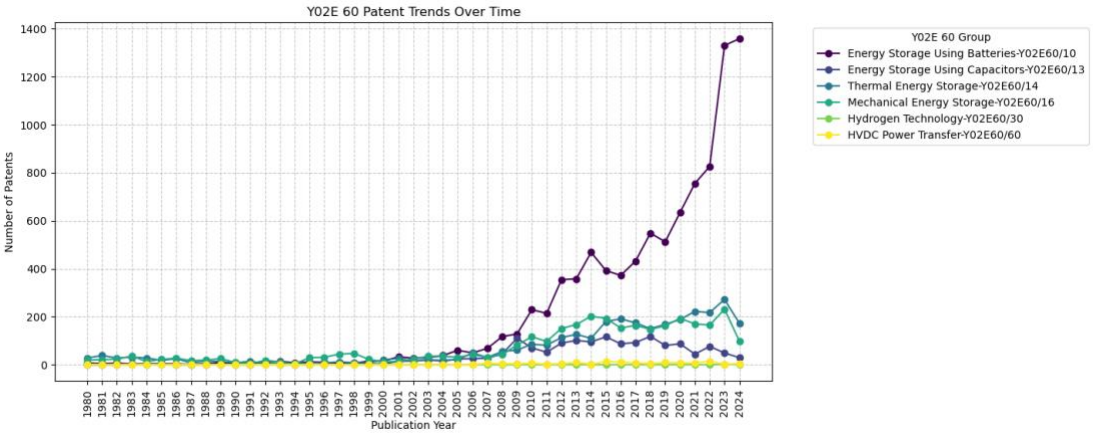
Note: *This figure displays the number of patents across specialized subgroups within Y02E10, highlighting the technical areas of highest innovation intensity. The most frequently patented technologies are wind turbines with rotation axis aligned to wind direction (Y02E10/72) and heat exchange systems (Y02E10/44), reflecting continued innovation in mechanical efficiency and energy capture. Other prominent areas include power conversion (MPPT) systems, solar mounting and tracking structures, and PV systems with concentrators. These results underline the technological diversity within the renewable energy domain, particularly in mechanical and photovoltaic integration approaches.*

Figure 20: *Patent Distribution by Y02E60 (Energy Storage Technologies).*



Note: *This chart illustrates the distribution of patents within the Y02E60 classification, which covers enabling technologies for energy storage. The dominant subgroup is Energy Storage Using Batteries (Y02E60/10), reflecting its central role in decarbonization and electrification. Other notable areas include mechanical and thermal energy storage solutions, which are gaining interest for grid stability and renewable energy integration. Capacitor-based storage, hydrogen, and HVDC technologies appear less frequently, suggesting either niche application areas or emerging innovation fronts within the energy transition landscape.*

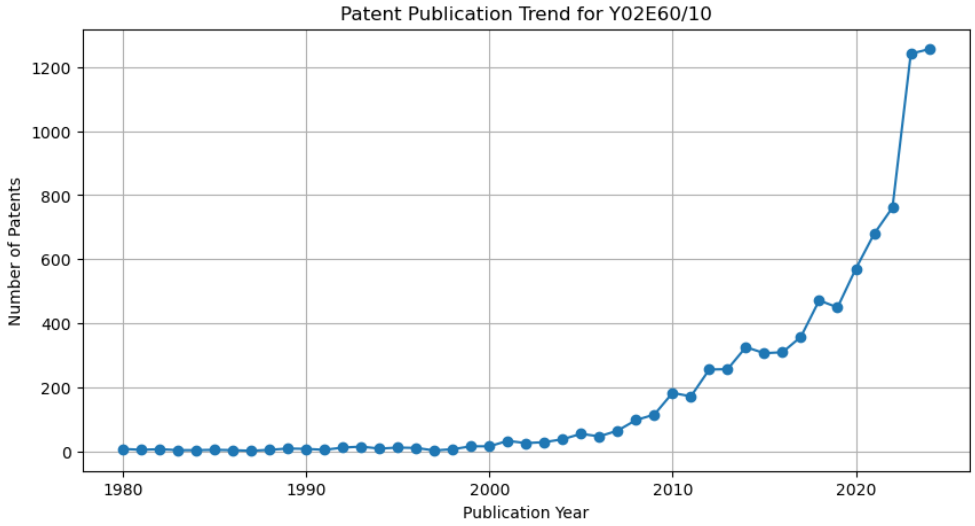
Figure 21: *Y02E60 Patent Trends Over Time by Subgroup*



Note: *This line chart displays the temporal trends in patent filings across key Y02E60 subgroups related to energy storage technologies. The data highlights a steep and sustained rise in patents for Energy Storage Using Batteries (Y02E60/10) since 2010, underscoring its dominant role in enabling renewable energy systems. Growth in mechanical and thermal energy storage is also evident, reflecting broader innovation efforts to diversify grid-scale storage solutions. By contrast, hydrogen technologies, capacitors, and HVDC systems show limited but steady activity, suggesting these areas remain either specialized or in earlier phases of technological development.*

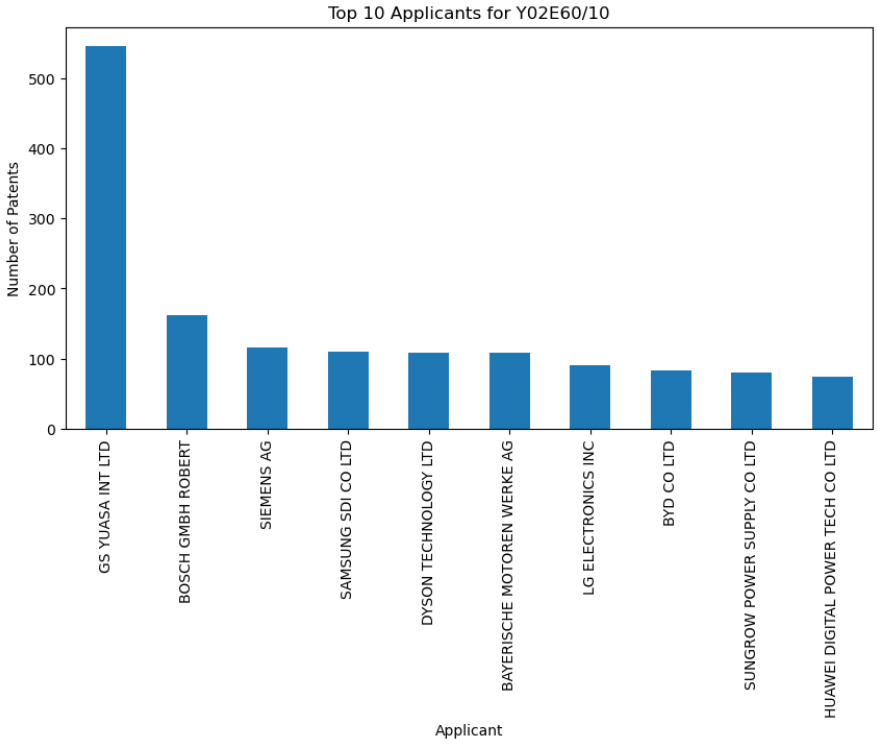
Appendix 2. Y02E60/10 Battery Deep Analysis

Figure 22: *Publication Trend for Battery Storage Technologies (Y02E60/10).*



Note: *This line graph shows the number of published patents over time within the CPC subclass Y02E60/10, which covers energy storage using batteries. The data reveals a slow but steady increase from 1980 until around 2008, followed by a dramatic acceleration through the 2010s and early 2020s. This surge reflects the global prioritization of battery innovation to meet growing demands for electric mobility, renewable energy integration, and sustainable energy storage solutions. The sharp rise in patent activity since 2020 underscores the strategic relevance of battery technology in the transition toward low-carbon energy systems.*

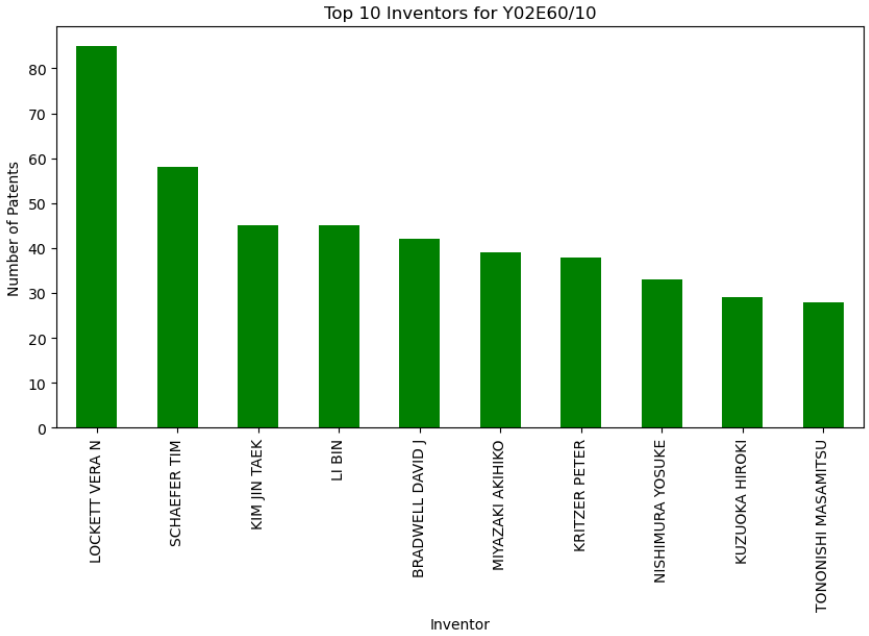
Figure 23: *Top 10 Patent Applicants in Energy Storage Using Batteries (Y02E60/10).*



Note: *This bar chart displays the leading applicants in the CPC subclass Y02E60/10, which focuses on battery-based energy storage technologies. GS Yuasa International Ltd. stands out with over 500 patent filings, significantly ahead of other top entities such as Bosch GmbH, Siemens AG, and Samsung SDI. The chart highlights the dominance of automotive and electronics companies in battery innovation, reflecting global investment in electric mobility, stationary storage, and renewable integration. The*

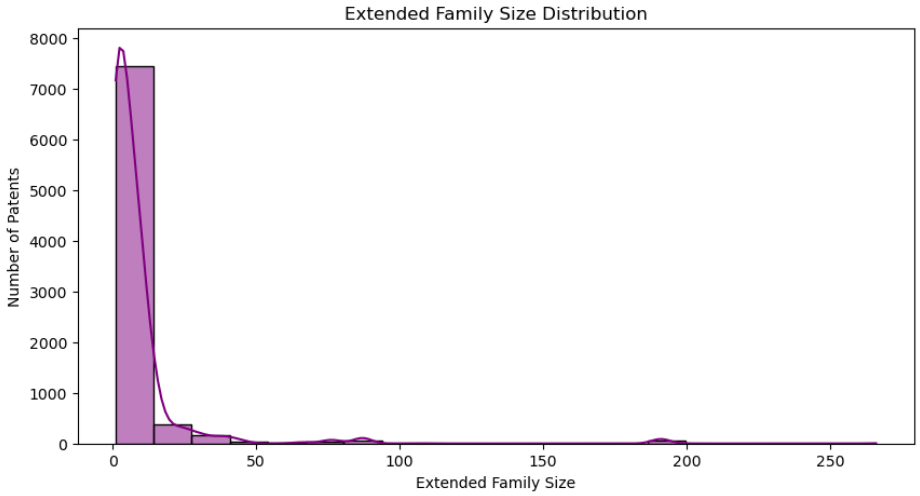
presence of both traditional manufacturers and tech firms underscores the strategic importance of battery IP across multiple sectors.

Figure 24: *Top 10 Inventors in Battery Storage Patents (Y02E60/10).*



Note: *This chart presents the most prolific individual inventors in the CPC subclass Y02E60/10, which focuses on battery-based energy storage technologies. Vera N. Lockett leads with over 85 patent filings, followed by Tim Schaefer and Jin Taek Kim. The list features inventors from both industrial and academic backgrounds, illustrating a broad base of innovation activity across geographical and institutional boundaries. The diversity in inventor affiliations emphasizes the global and interdisciplinary nature of battery technology development.*

Figure 25: *Citation and Family Size Distributions for Y02E60/10 Patents.*

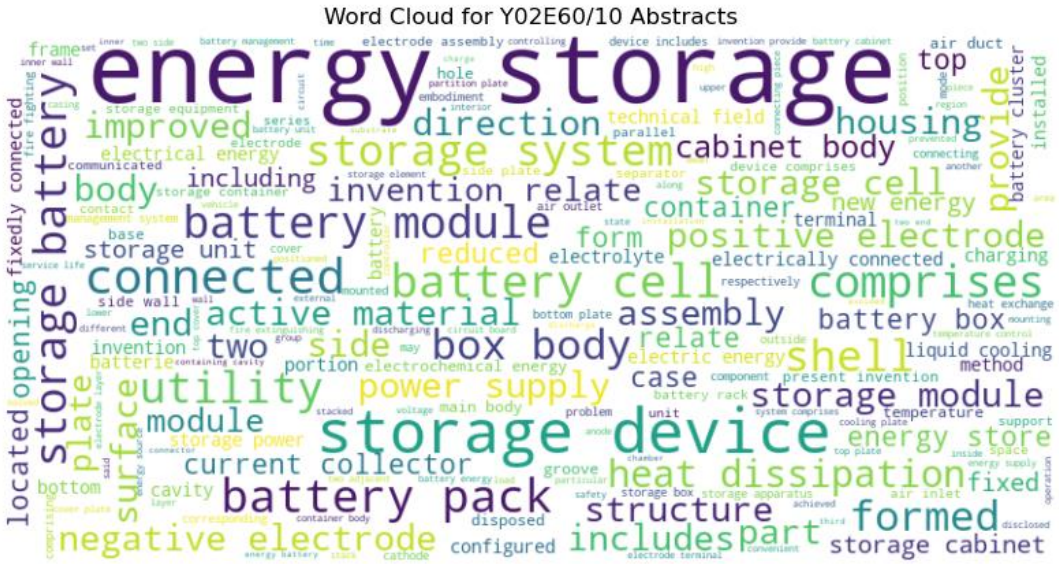


Note: *The upper panel illustrates the distribution of forward citation counts, showing that most patents within the Y02E60/10 subclass receive relatively few citations, with a small subset achieving exceptionally high citation counts—indicative of breakthrough or foundational innovations. The lower panel presents the distribution of extended patent family sizes, which reflect the geographic and strategic breadth of patent protection. Similar to citation patterns, most patents belong to smaller families, while a few are part of significantly larger international portfolios, suggesting their*

commercial or strategic importance. Both distributions highlight the skewed nature of innovation impact in battery technologies.

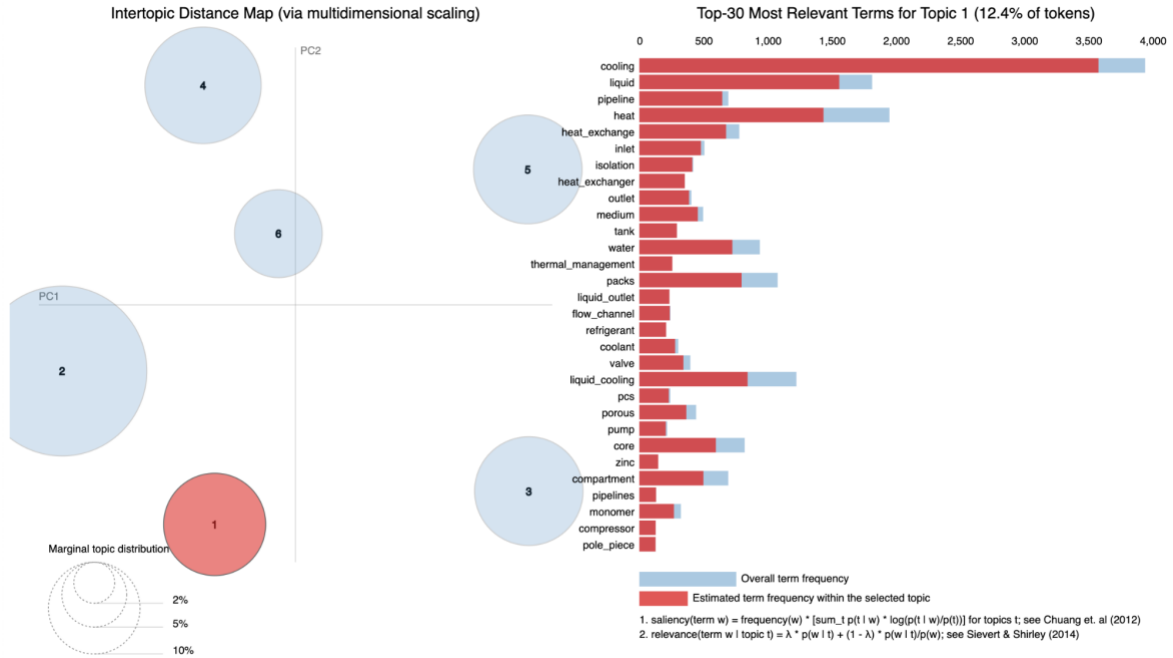
Appendix 3. Natural Programming Language (NPL) Analysis.

Figure 26: *Most Frequent Terms in Patent Abstracts (Y02E60/10).*



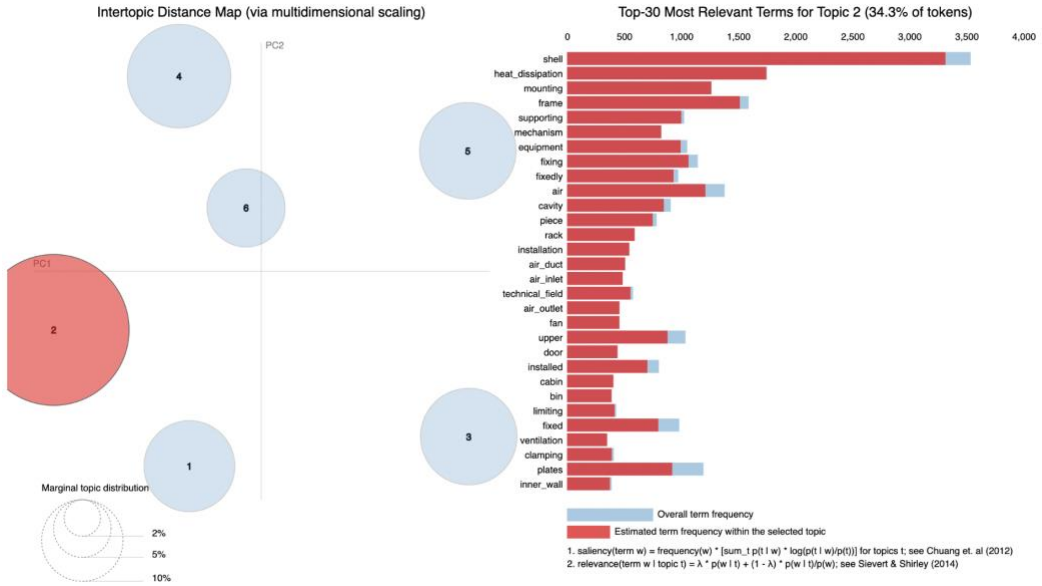
Note: *This word cloud visualizes the most frequent terms found in the abstracts of patents classified under CPC Y02E60/10, which focuses on battery-related energy storage technologies. Larger words represent higher frequency, highlighting key technical terms such as "energy," "storage," "battery," "module," and "electrode." The prominence of terms like "heat dissipation," "shell," and "cooling" suggests a strong emphasis on thermal management and structural design, which aligns with the dominant topics identified through LDA topic modeling. This visualization provides an accessible summary of the thematic focus areas in the patent corpus.*

Figure 27: *Intertopic Distance Map and Top-30 Relevant Terms for Topic 1 (LDAvis Output).*



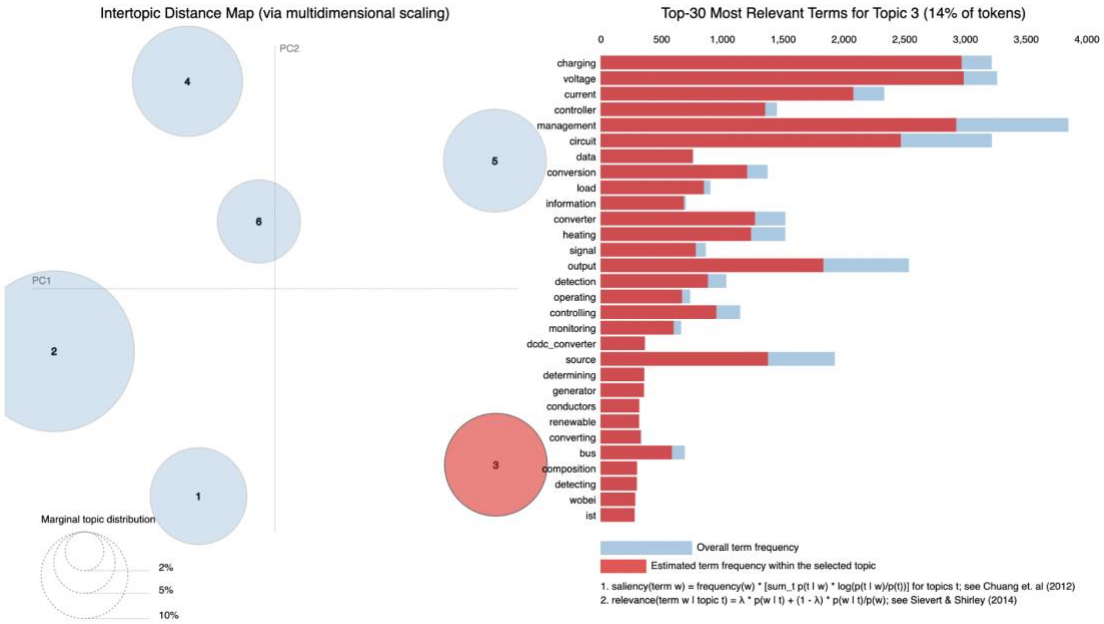
Note: *The left panel shows the intertopic distance map, where Topic 1 is visualized as a red bubble representing 12.4% of all tokens in the Y02E60/10 patent corpus. The size indicates relative topic prevalence, and the distance from other topics reflects semantic uniqueness. Topic 1 is positioned distinctly from Topic 2 and others, indicating a well-separated thematic focus on thermal systems.The right panel lists the 30 most relevant keywords for Topic 1 using a relevance parameter λ = 0.6. Keywords such as "cooling," "liquid," "pipeline," "heat exchanger," and "thermal management" clearly point to liquid-based cooling technologies used in battery systems. These results confirm the coherence and interpretability of the topic, aligning with known industrial efforts to improve heat regulation in high-performance battery modules (Sievert & Shirley, 2014).*

Figure 28: *Intertopic Distance Map and Top-30 Relevant Terms for Topic 2 (LDAvis Output).*



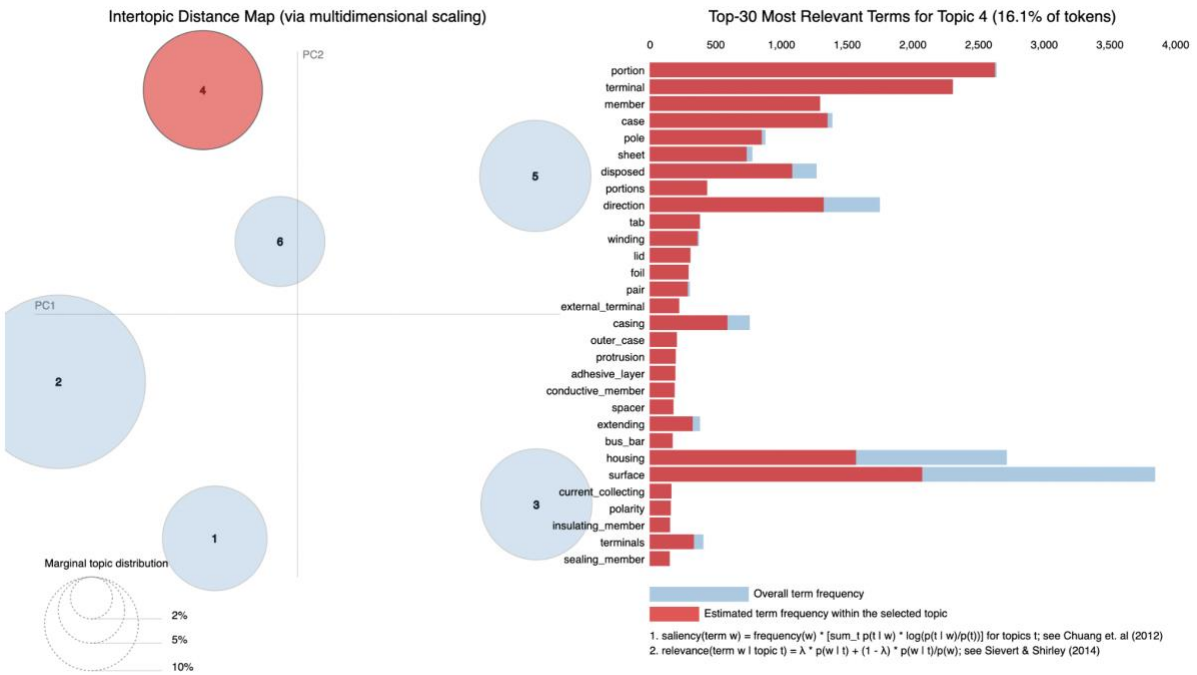
Note: *The left panel shows the intertopic distance map, generated using multidimensional scaling (MDS), where each bubble represents a topic identified by the LDA model applied to Y02E60/10 battery patents. The size of each bubble indicates the prevalence of the topic in the corpus, with Topic 2 (in red) dominating the distribution at 34.3% of all tokens. The distance between bubbles reflects semantic similarity—closer bubbles share more overlapping terms. The right panel displays the top 30 most relevant terms for Topic 2, which relates to battery enclosures and heat dissipation. Red bars represent the estimated frequency of each term within the topic, while blue bars indicate overall frequency across the entire dataset. Terms like "shell," "heat_dissipation," and "mounting" suggest a strong thematic focus on mechanical structures designed for cooling and support in battery systems. The λ = 0.6 relevance setting balances specificity and saliency of terms (Sievert & Shirley, 2014).*

Figure 29:. *Intertopic Distance Map and Top-30 Relevant Terms for Topic 3 (LDAvis Output).*



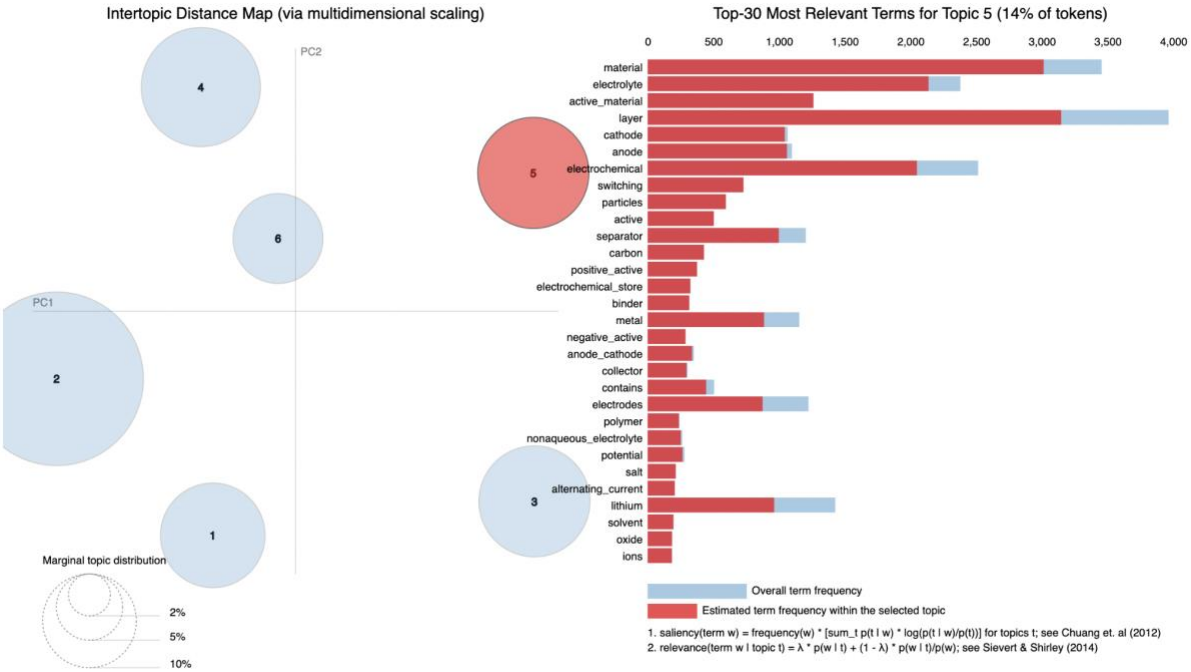
Note: *The left panel displays the intertopic distance map, with Topic 3 shown as a red bubble accounting for 14% of all tokens in the Y02E60/10 corpus. It is semantically distinct from other topics, reinforcing the thematic clarity of this subject area. The right panel ranks the 30 most relevant terms for Topic 3, using a relevance setting of λ = 0.6. Key terms include "charging," "voltage," "current," "controller," and "dc/dc converter," which suggest a thematic focus on battery energy management, power regulation, and smart charging systems. The separation and coherence of Topic 3 validate its importance in addressing emerging demands such as fast charging, electric vehicle grid integration, and intelligent power control architectures (Sievert & Shirley, 2014).*

Figure 30: *Intertopic Distance Map and Top-30 Relevant Terms for Topic 4 (LDAvis Output).*



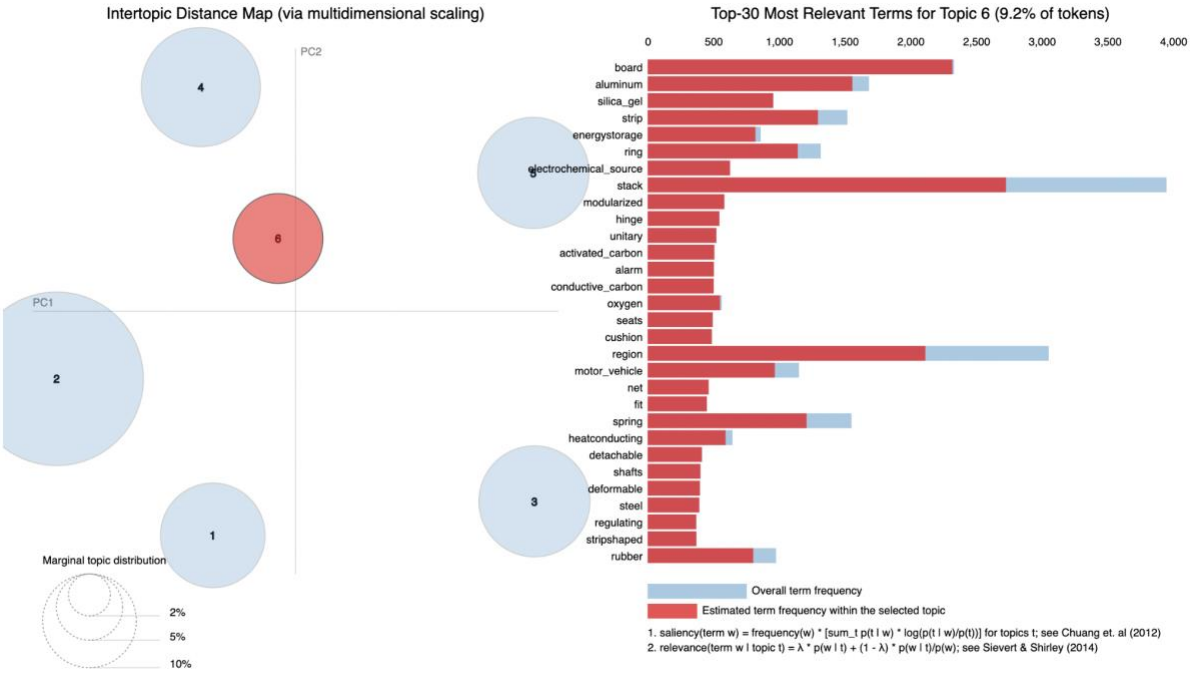
Note: *The left panel shows the intertopic distance map, where Topic 4 appears as a clearly defined red bubble, accounting for 16.1% of the token distribution in the Y02E60/10 dataset. Its spatial separation from other topics indicates thematic distinctiveness. On the right, the bar chart displays the top 30 terms most strongly associated with Topic 4, based on a relevance metric (λ = 0.6). Notable terms such as "terminal," "housing," "bus_bar," "sealing_member," and "current_collecting" suggest that Topic 4 centers on the physical and structural components of battery modules. This includes innovations in external housing, electrical connections, and insulation, all of which are critical for improving safety, durability, and integration of battery systems into larger energy storage infrastructures.*

Figure 31: *Intertopic Distance Map and Top-30 Relevant Terms for Topic 5 (LDAvis Output).*



Note: *The left panel displays the intertopic distance map, where Topic 5 is positioned distinctly, suggesting it captures a unique theme within the corpus. It represents 14% of the token distribution. The right panel shows the 30 most relevant terms for Topic 5, indicating a clear focus on electrochemical layer materials. Key terms such as "active_material," "electrolyte," "cathode," "anode," "separator," and "electrochemical" highlight innovations related to battery chemistry and core functional materials. These materials are central to energy density, charge/discharge efficiency, and lifecycle performance in modern energy storage technologies (Nanda et al., 2021; Zhang et al., 2022). The distinctiveness and technical specificity of this topic reinforce the LDA model's ability to differentiate material-level research within battery innovation.*

Figure 32: *Intertopic Distance Map and Top-30 Relevant Terms for Topic 6 (LDAvis Output).*



Note: *The intertopic distance map (left) positions Topic 6 as relatively distinct and compact, representing 9.2% of the token distribution. The right panel highlights the 30 most relevant terms, prominently featuring keywords like "stack," "modularized," "strip," "detachable," "conductive_carbon," and "motor_vehicle." These suggest a focus on battery stack design and protection modules, with attention to structural integration, modular assembly, and thermal/mechanical resilience—especially relevant in automotive or mobile applications. The emergence of terms like "spring," "hinge," and "rubber" underscores the engineering of flexible or shock-resistant systems, validating the LDA model's ability to isolate form-factor and packaging innovations within battery technology.*