data

MDPI

Article

Natural Language Processing Patents Landscape Analysis

Hend S. Al-Khalifa [1,2,*], Taif AlOmar [2] and Ghala AlOlyyan [2]

[1] Department of Information Technology, College of Computer and Information Sciences, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia
[2] iWAN Research Group, King Saud University, Riyadh 11543, Saudi Arabia
[*] Correspondence: hendk@ksu.edu.sa

Abstract: Understanding NLP patents provides valuable insights into innovation trends and competitive dynamics in artificial intelligence. This study uses the Lens patent database to investigate the landscape of NLP patents. The overall patent output in the NLP field on a global scale has exhibited a rapid growth over the past decade, indicating rising research and commercial interests in applying NLP techniques. By analyzing patent assignees, technology categories, and geographic distribution, we identify leading innovators as well as research hotspots in applying NLP. The patent landscape reflects intensifying competition between technology giants and research institutions. This research aims to synthesize key patterns and developments in NLP innovation revealed through patent data analysis, highlighting implications for firms and policymakers. A detailed understanding of NLP patenting activity can inform intellectual property strategy and technology investment decisions in this burgeoning AI domain.

Keywords: NLP; patents; innovation; artificial intelligence; lens patent database

Citation: Al-Khalifa, H.S.; AlOmar, T.; AlOlyyan, G. Natural Language Processing Patents Landscape Analysis. *Data* **2024**, *9*, 52. https://doi.org/10.3390/data9040052

Academic Editors: Francesco M. Donini and Joaquín Torres-Sospedra

Received: 14 November 2023
Revised: 11 January 2024
Accepted: 28 March 2024
Published: 31 March 2024

1. Introduction

Natural language processing (NLP) is a field of computer science that focuses on the interaction between computers and human (natural) languages. It is a sub-field of artificial intelligence (AI) that deals with the ability of computers to understand and process human language, including speech and text.

NLP is used in a variety of applications, including machine translation, speech recognition, text analysis, and question answering. It is also used in chatbots and virtual assistants, such as Siri and Alexa. NLP is a complex and challenging field, but it is also a very promising one. As technology continues to advance, NLP is likely to become even more powerful and widespread.

The market for natural language processing (NLP) is expected to grow rapidly in the coming years. According to a market report by Statista [1], the NLP market size is projected to reach US$1.58 billion in 2023 and is projected to show an annual growth rate (CAGR) of 14.81% during the forecast period (2023–2030). This growth is being driven by the increasing use of NLP in a variety of applications, such as chatbots, machine translation, and voice recognition. The growth of the NLP market is being supported by a number of factors, including the increasing availability of data, the development of new algorithms, and the growing demand for NLP-powered applications. These factors has made the landscape for NLP patents proliferate rapidly [1].

NLP technologies power many AI systems we use every day, like Siri, Alexa and ChatGPT. However, research on NLP patents specifically is lacking. Haney [2] provided the first in-depth analysis of NLP patents, discussing the main technical approaches to NLP software and modeling trends in the NLP patent market. However, more comprehensive surveys of the NLP patent landscape are needed to provide researchers and companies with key strategic insights to guide innovation and business decisions. Thorough analysis of the NLP patent space can inform research and development (R and D) planning, Intellectual

Property (IP) strategy, partnerships and competitive benchmarking for researchers and businesses by identifying crucial information on competitors, technologies, and opportunities. Therefore, this research seeks to answer the following questions:

1. What is the temporal trend in NLP patent publications? And languages?
2. Which jurisdictions have the most NLP patents?
3. Who are the top applicants and inventors in NLP patents?
4. What are the most common classifications for NLP patents?
5. What are the important NLP patents?
6. How have the abstracts of NLP patents evolved over time? Can we identify any trends or shifts in focus?
7. What is the average time from application to publication for NLP patents?
8. What is the legal status of these patents? Are there any patterns or trends?

This analysis contributes to the research community in several ways, particularly for researchers in the field of Natural Language Processing (NLP) and those interested in intellectual property and innovation studies. Here are some key contributions:

1. Understanding the NLP Patent Landscape: This analysis provides an overview of the patent landscape in NLP, including trends in patent applications over time, the jurisdictions where patents are filed, and the key players in terms of inventors and assignees. This information can help researchers understand where and how innovation is happening in the NLP field.
2. Insights into Patent Characteristics: The analysis of patent duration, legal status, and patterns provides insights into the characteristics of NLP patents. These insights can help researchers understand the patenting process in the NLP field and the strategies used by inventors and assignees.
3. Data for Further Research: The cleaned and processed dataset used in this analysis can serve as a valuable data source for further research. For example, researchers could use this data to conduct more in-depth studies of NLP patents, such as citation analysis, network analysis, or text mining of patent abstracts.
4. Implications for Policy and Practice: The findings from this analysis could inform policy discussions about innovation in the NLP field, as well as practical decisions by inventors, companies, and other stakeholders about patent strategy.

The rest of the paper is organized as follows. Section 2 presents previous related work in the area of AI and NLP patents. Section 3 goes over the research questions and provides answers to them. Finally, Section 4 concludes the paper with important recommendations and future outlooks.

2. Related Work

Artificial intelligence is rapidly advancing and transforming numerous industries. As such, AI has become a hot topic in both technology and law. Several studies have analyzed how AI intersects with and challenges intellectual property law, particularly the patent system.

Warin et al. [3] used social data science techniques to analyze 55,109 AI patents and 29,225 related articles. The study found that AI patents have grown exponentially over time and cover a wide range of sub-fields. Yoo et al. [4] proposed a method for analyzing AI patents using vector space models and deep learning models, like KeyBERT. A case study shows how the model can extract keywords and analyze relationships between AI technologies.

Ngoc and Ngoc [5] provided a comprehensive analysis of AI patents in the U.S. and Vietnam. The study found that the U.S. leads in AI research and patent filings but questions whether and how the U.S. protects AI inventions. The paper also examines liability issues with AI applications. Haney [6] introduced an original dataset of four types of machine learning patents: deep learning, reinforcement learning, deep reinforcement learning, and natural language processing. Analysis of the dataset provides insights into the AI patent landscape and reveals significant overlap between technologies. Leusin et al. [7] presented a study of patenting patterns in AI across different countries and techniques. The authors propose two novel indicators to measure the national and international attractiveness of countries for AI development and protection. They use patent data and keywords to identify AI-related patents and analyze their trends over time. They found that China and the US are the dominant national breeding grounds for AI, while the US and some European countries are the major international breeding grounds. They also observe significant changes in the technological leadership and the relevance of AI techniques over time.

When it comes to NLP patent analysis, Chao et al. in [8] provided a comprehensive study of the emerging technologies in the field of NLP with an intense focus on exploring the technological developmental trends of NLP-enabled intelligent chatbots. The study adopted several text-mining techniques, such as document term frequency analysis (TF-IDF) to identify the key terminologies used in NLP patents, clustering methods to analyze the sub-domains of NLP that have a higher application rate in patents, and Latent Dirichlet Allocation for topic modeling to extract the key topics in the patent dataset. The paper found that, according to their analysis of the Derwent Innovation dataset, the patent exploiting NLP technologies started in 2014 and developed rapidly until 2019, with a large number of applications in speech recognition following the research maturity of the domain in 2018, since when several speech recognition techniques have been developed and perfected. The paper also proposed a systematic domain-agonistic patent analysis methodology that can be used by researchers to examine the trends and emerging technologies in different fields.

Moreover, other than the paper mentioned above, we did not find many research papers that fulfill the goal of analyzing the NLP patent landscape; instead we found research in which NLP can be used to analyze patents themselves. For instance, Arts et al. [9] developed NLP techniques to identify the creation of new technologies in patents. They validate their techniques using patents linked to major awards, which are likely to cover radically new technologies, and patents granted in the U.S. but rejected elsewhere, which are likely to lack novelty. They provide open-source code and data for analyzing all U.S. utility patents up to 2018. Similarly, Balsmeier et al. [10] applied machine learning and NLP to build a database of inventors, assignees, and locations mentioned in U.S. patents from 1976 to 2016. They introduce a novelty measure based on first word usage and tools for mapping co-inventor networks and visualizing patenting trends.

In summary, while research on NLP patents is limited, the available papers show that NLP can be used to analyze patents, extract knowledge from them, identify new technologies within them, and build useful databases and tools for working with patent data. However, in this research we will analyze the NLP patent landscape in order to understand its patterns and trends.

3. Data and Method

To analyze the developmental evolution of NLP technologies, trace their potential pathways, and answer our research questions, we collected a dataset of published patents in the NLP domain. The Lens [11] database was used as the main platform for the data collection process as it contains patents granted from different patent offices across the world; thus, it does not restrict the search to a particular patent office or jurisdiction. Moreover, Lens allows users to execute advanced search queries on its database, which simplifies the process of collecting all related patents and minimizes the need for manual data cleaning and inspection. The main limitation of Lens is that it offers access to only the first 50,000 patents from the search results; however, we were able to bypass this limitation by grouping the keywords such that the search result of each group does not exceed 50,000 patents; then we aggregated the results of each group and dropped any duplicated patents. This retrieval process resulted in a dataset of 302,934 patents, in which 99,857 are active.

The following is a list of all the keywords used in the data retrieval process: “natural language processing”, “natural language understanding”, “human language processing”, “human language understanding”, “NLP algorithms”, “semantic analysis”, “lexical analysis”, “language translation”, “machine translation”, “neural networks and language processing”, “speech recognition”, “voice recognition”, “text-to-speech”, “speech-to-text”, “chatbot”, “sentiment analysis”, “natural language”, “language processing”, “language understanding”, “G06F40/00” (IPC Class).

The last keyword “G06F40/00” is an International Patent Classification (IPC) class that contains patents concerning “Handling natural language data”. This class was explicitly chosen because it is closely related to the NLP domain. Finally, the keywords search was conducted on the title, abstract and claims of each patent using the “OR” operator.

To better understand our dataset, Figure 1 presents the outcomes of applying Bertopic [12] to identify the most pertinent topics within the dataset by clustering patents that share similarities or revolve around the same technological domain. Bertopic was employed on the title of each patent after pre-processing, which involved eliminating stop words and punctuation marks.

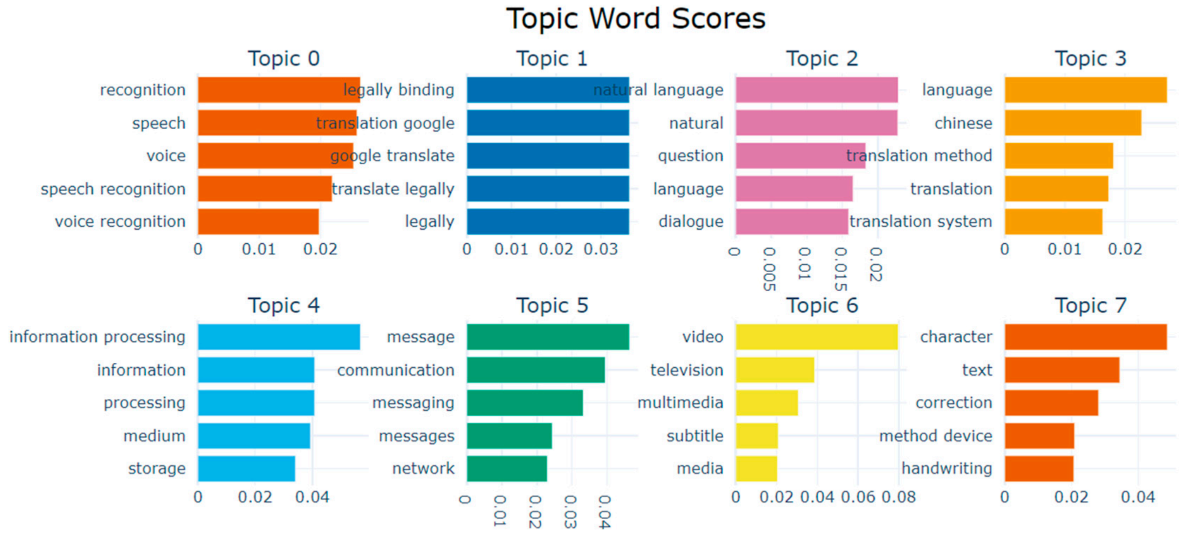

Figure 1. Top 7 topic clusters formed using Bertopic to cluster the patents based on their focus.

Regarding Bertopic hyper-parameters, the minimum topic size was established at 50 to prevent the formation of small clusters and to reduce the overall number of clusters. Additionally, the n-gram model range was set to (1,2), since most NLP topics typically consist of two words, such as machine translation, speech recognition, and language processing.

According to Figure 1, most patents in the dataset are related to speech recognition, since this is the largest cluster formed, with 39,398 patents. The second and fourth largest clusters contain patents filed for language translation, especially Google Translate.

4. Analysis and Discussion

This section will answer the posed research questions and discuss their results.

4.1. NLP Patent Publications Trend and Language

Figure 2 shows the temporal trend in NLP patent publications. This trend depicts the importance of NLP over time by visualizing the number of patents per year starting from the early 1900s. The publication line appears to increase over time, indicating a growing trend in the number of NLP patent publications, specifically after 2010. Moreover, the highest increase in the number of patents filed can be observed between 2018 and 2022. This peak might be due to the sudden increase in advancements in technology, increased funding, or a growing recognition of the potential applications of NLP, such as Large Language Models (LLMs). On the other hand, we might observe a drop or plateau in the most recent years. This could be due to incomplete data for the current year, or it might suggest a slowdown in the publication of new patents.

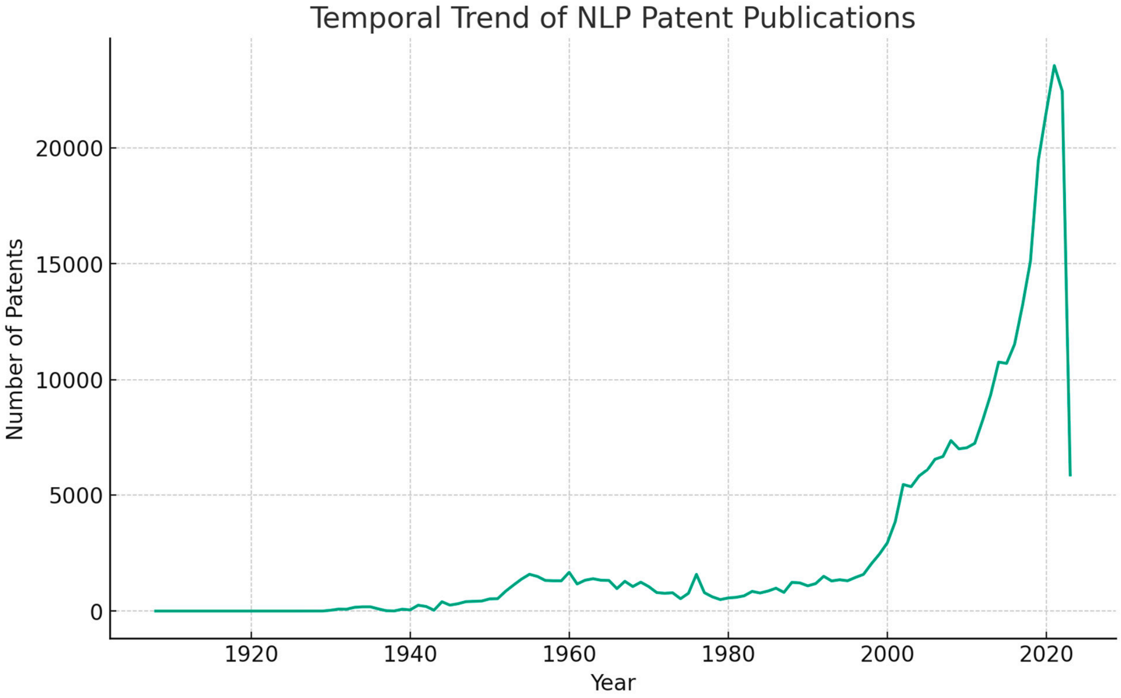

Figure 2. Distribution of patents in the field of NLP over the years.

It is worth mentioning that the oldest patent in this dataset was released on 24 September 1908. This patent, identified by the display key 'GB 190812735 A', bears the title *An Improved Book, Pamphlet, or Similar Item for Enhancing Conversation in Foreign Languages*. Charles Hugo and Stephen Armstrong are the applicants credited for this patent.

As for the language that the NLP patent is addressing, we extracted each language from our collected dataset. In total, there are 53 languages mentioned in the titles and abstracts, where Table 1 shows the top 10 languages in terms of occurrence. We can observe that Chinese came first with a count of 3489 patents, and English in second place with 2286 patents. Although we did not see Japanese NLP patents in recent years, it achieved third place due to its continuous prosperity between 1985 and 2007.

Table 1. Distribution of Languages in the NLP patent dataset.

Language	**Count**
Chinese	3489
English	2286
Japanese	1439
Korean	375
French	168
German	147
Arabic	143
Vietnamese	97
Latin	80
Tibetan	71

4.2. NLP Patent Jurisdictions

The geographical distribution of NLP patents indicates which regions lead in NLP research. As we can see from Figure 3, we can say that the United States (US) and China (CN) clearly dominate in terms of the number of NLP patents. This could be an indicator of the significant focus on technological advancement and investment in NLP research in these two countries. Spain (ES) also shows a high number of NLP patents, suggesting significant activity in this field.

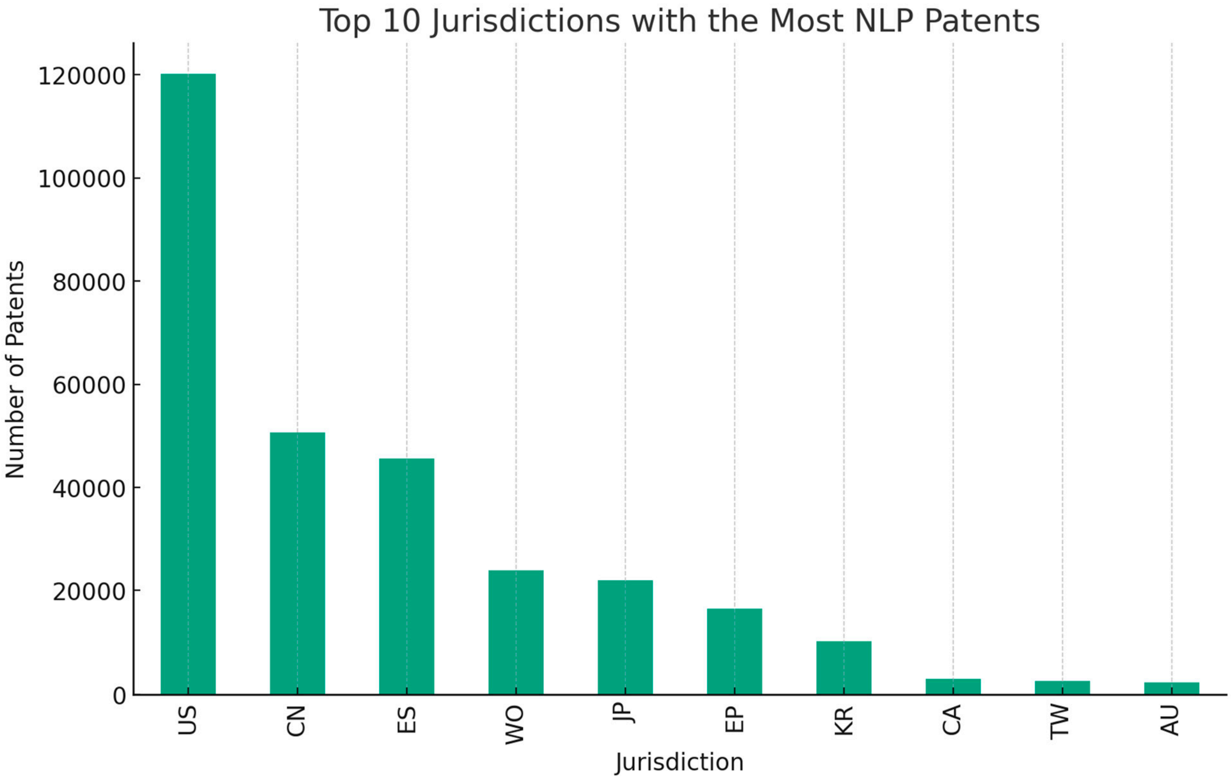

Figure 3. Top 10 jurisdictions with the most NLP patents.

The World Intellectual Property Organization (WIPO), representing international patents, also has a significant number of NLP patents. This suggests that many inventions in the field of NLP are intended for global use, hence the international patents.

Other jurisdictions with substantial numbers of NLP patents include Japan (JP), the European Patent Office (EP), and South Korea (KR). This indicates that these regions also have significant activity in NLP research and development.

Canada (CA), Taiwan (TW), and Australia (AU) have comparatively fewer NLP patents. This could be due to various reasons, such as the size of the tech industry, the focus of technological development, or the overall patenting activity in these regions.

4.3. Top Applicants and Inventors

This section highlights the companies and universities at the forefront of NLP patenting.

Figure 4 provides a visual representation of the top 10 applicants for NLP patents. From the graph, we can observe that IBM is the leading applicant for NLP patents, with 16,103 patents. Likewise, Microsoft follows closely behind as the second most prolific applicant, providing 11,077 patents, which serves as evidence of their dedicated efforts in developing NLP technologies. Google secures the third place with 6033 patents, slightly surpassing Samsung Electronics. The graph shows the presence of both Western and Eastern companies, demonstrating that NLP research and innovation is a global endeavor. Finally, these top 10 companies emerge as the primary contributors to the NLP field.

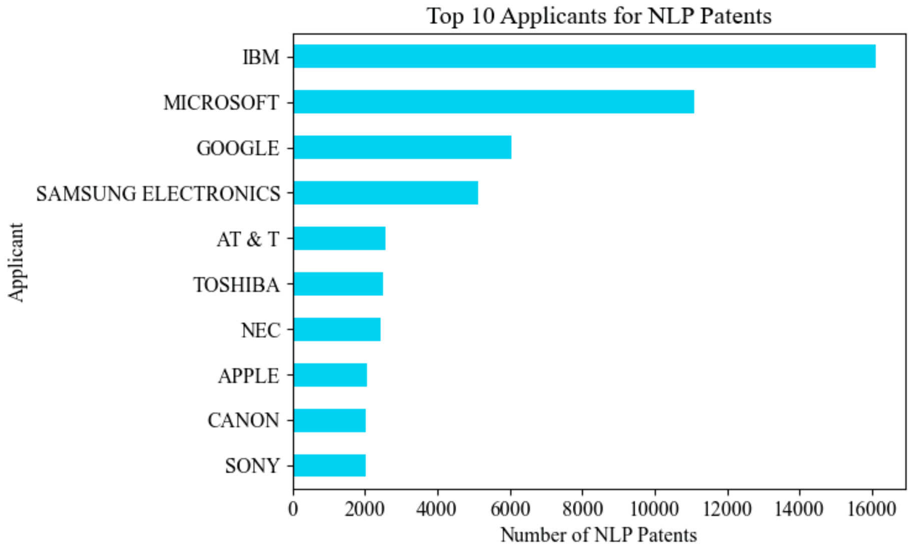

Figure 4. Top 10 applicants for NLP patents.

From an academic standpoint, Figure 5 showcases the leading 10 educational institutions that submitted patents related to Natural Language Processing (NLP). The Institute of Computing Tech CN Academy leads with 44 patents. CollegeNET Inc follows with 16 patents. The University of Science and Technology, Beijing, Chongqing College of Electric Engineering, and the People's Liberation Army National University of Defense Technology each have 14 patents. The Aerospace Information Research Institute, Chinese Academy of Sciences, has applied for 13 patents, and the Alibaba DAMO Academy Hangzhou Tech Co., Ltd. possesses 12 patents. The remaining three institutes each have 9 patents.

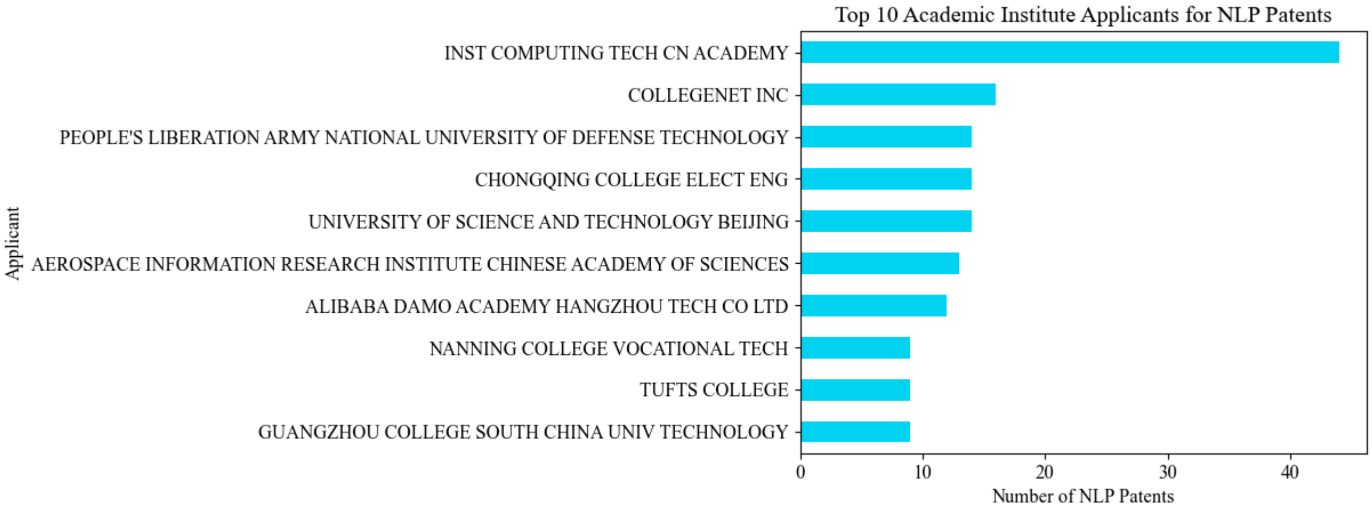

Figure 5. Top 10 academic institute applicants for NLP patents.

The graph also reveals a significant difference in patent volumes between the top academic institutions and industry players. In the industry sector, the leaders in patent applications, such as IBM, Microsoft, and Google, are U.S.-based, reflecting the significant role of American technology companies in global innovation. These companies have substantial resources and a strong commercial incentive to invest in research and development, securing patents to protect their investments and maintain a competitive edge.

On the other hand, the academic sector is led by Chinese institutions, such as the Institute of Computing Tech CN Academy. This demonstrates the substantial investment by China in academic research and technology. It is a testament to China's strategic focus on developing its domestic technology capabilities and underscores the global nature of research and innovation.

Table 2 illustrates the top 10 inventors ranked by the number of patents held, along with their corresponding patent filing jurisdictions. Leading the list is Vadim Fux with 413 patents filed under the World Intellectual Property Organization (WIPO) jurisdiction, signifying an international scope. He is followed closely by Wang Jianzong with 354 patents, also filed under WO. Inventors Rakshit Sarbajit K, Acero Alejandro, and Allen Corville O, each with over 300 patents, represent the United States in this list, highlighting the strong innovation ecosystem in the US. Wu Hua stands out as the only inventor in the top 10 with patents filed under the European Patent Office (EP) jurisdiction. Wang Haifeng and Li Wei, like the top two inventors, have their patents filed internationally (WO), while Trim Craig M and Gillick Dan, both with over 260 patents, add to the US representation. Overall, this data underscores the global nature of innovation, the significant role of individual inventors in advancing technology, and the prominent position of the US in patent activity.

Table 2. Top 10 inventors based on number of patents, along with their corresponding jurisdiction of filing.

Inventor	**Number of Patents**	**Jurisdiction**
Fux, Vadim	413	WO
Wang Jianzong	354	WO
Rakshit Sarbajit K	320	US
Acero Alejandro	310	US
Allen Corville O	309	US
Wu Hua	293	EP
Wang Haifeng	276	WO
Li Wei	272	WO
Trim Craig M	267	US
Gillick Dan	266	US

4.4. Common Classifications for NLP Patents

International Patent Classification—Reformed (IPCR) is a hierarchical system for classifying patents based on the technological areas they target. The IPCR is the reformed version of the IPC, International Patent Classification system established in 1971. Figure 6 demonstrates the top 10 most frequent IPCR classes found in our dataset:

G06F17/30: Information retrieval; database structures; file system structures.

G10L15/22: Procedures used during a speech recognition process, e.g., man-machine dialog.

G06F40/00: Handling natural language data.

G10L15/26: Speech-to-text systems.

G06F17/27: Digital computing or data processing equipment or methods, specially adapted for specific functions, Automatic analysis, e.g., parsing, orthograph correction.

G06F17/28: Processing or translating of natural language.

G10L15/00: Speech recognition.

G06F3/16: Sound input; sound output.

G10L15/06: Creation of reference templates; training of speech recognition systems, e.g., adaptation to the characteristics of the speaker’s voice

G10L15/18: Speech recognition using natural language modeling.

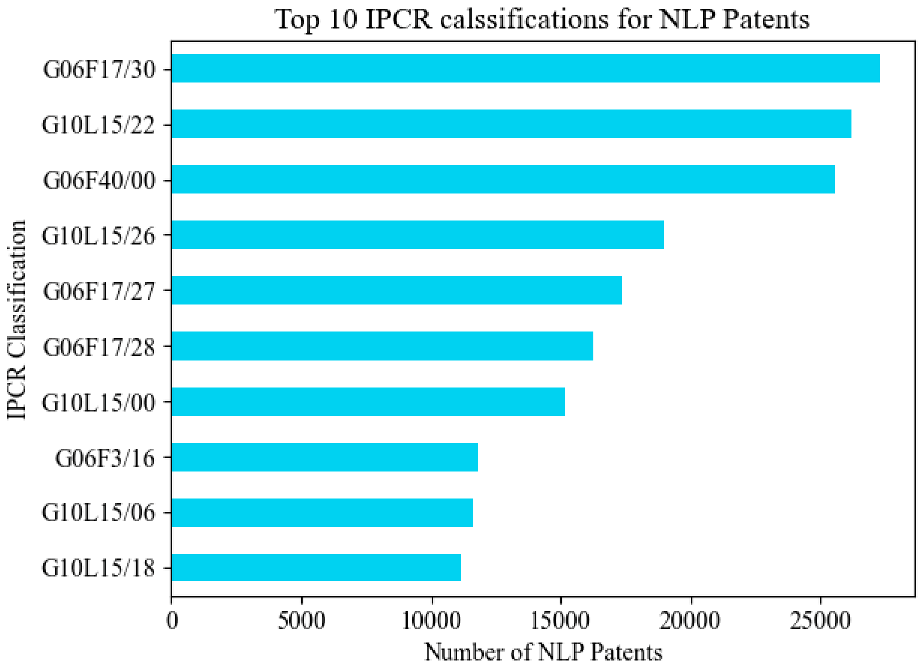

Figure 6. Top 10 IPCR (International Patent Classification—Reformed) classifications for NLP patents.

As can be seen in Figure 6, around 27,000 patents, which make up 9% of our dataset, are filed under the IPCR class G06F17/30, which describes patents in the field of information retrieval and database structure. This indicates that a good portion of the patents related to NLP focus on describing methodologies for accessing and obtaining the most appropriate data from repositories based on user-specified queries.

The second most prevalent IPCR class in our dataset is G10L15/22, with about 26,000 patents, which is intended for patents filed for speech recognition procedures and processes. This suggests a strong emphasis on voice-based technology in NLP patents, highlighting the importance of inventing new procedures of speech recognition and analysis in current technological advancements. Moreover, the increasing interest in automated personal assistants could also explain the high number of patents filed under this class.

Furthermore, the third most common IPCR class is G06F40/00, which covers 8.4% of our dataset. This class generally describes patents intended for handling natural language data. This is a very broad class that includes patents for text processing techniques, semantic analysis, and language translation; this explains why it is among the top three most frequent IPCR classes in our dataset. Moreover, G06F40/00 was used as a keyword during the dataset curation phase, as mentioned in the data and method section, which may have contributed to increasing the number of patents filed under this class.

Notably, the presence of G06F17/27 and G06F17/28, categorized under 'Electric Digital Data Processing', shows the intersection of NLP with broader data processing techniques. This points to the multidisciplinary nature of NLP innovations, where speech technology interfaces with other data processing methods, thereby expanding the potential of digital technologies.

Figure 6 also highlights the dominance of the general IPC classes G10L15/00 'Speech Recognition' and G06F17/00 'Digital Computing' in patent publications, since most of the IPC sub-classes shown in the figure fall under these categories. This emphasizes the importance, wide-use, and impact of speech recognition and data processing applications in the field of NLP.

In summary, these results provide an overview of the main areas of technology that NLP patents are focusing on. They suggest that information retrieval, speech recognition, speech-to-text systems, and natural language data processing are key areas of interest in NLP research and development.

4.5. Important NLP Patents

Inventions that are more important and relevant tend to have a higher number of citations; these inventions are usually the foundation on which other inventions and applications are built. As shown in Figure 7, the most cited patent in our dataset is "US 30933694 A" with 2562 citations. This patent is a US patent application granted in 20/04/1998 for "Personal Communications Internetworking". The second and third most cited patents are "US 54077200 A" filed for "Web client-server system and method for incompatible page markup and presentation languages", and "US 72231496 A" filed for "Computer-based communication system and method using metadata defining a control structure", respectively, each with citation counts exceeding 2000. All of these patents focus on data transfer methods through a communication network. They are not directly related to NLP but rather used in several NLP applications as a basis on which to establish data transfer and communication.

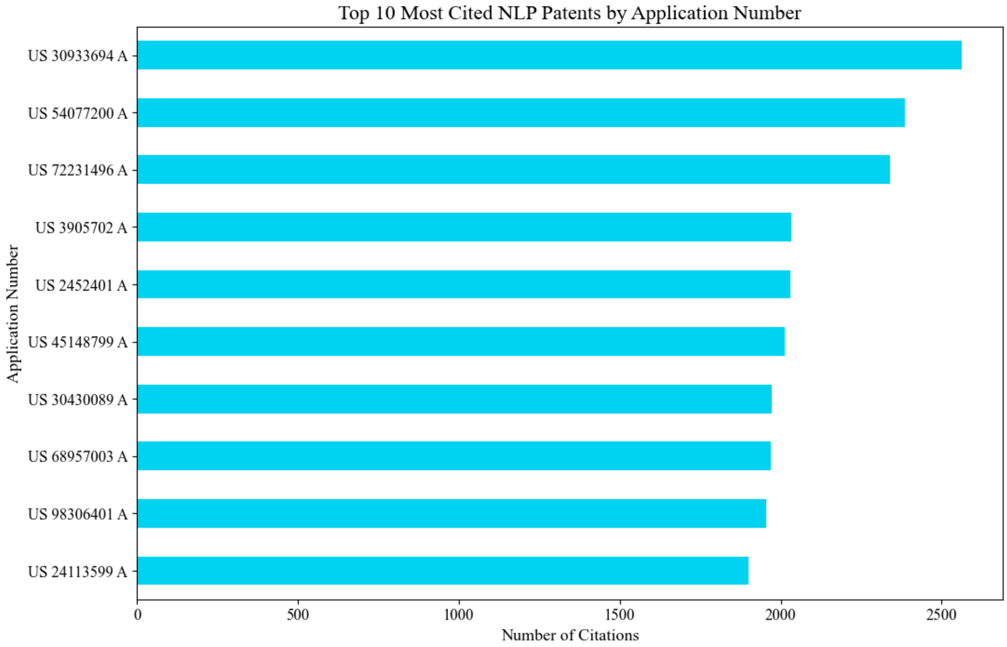

Figure 7. Top 10 most cited NLP patents, identified by their Application Numbers.

The remaining patents on the list, though having fewer citations, are still highly influential in the NLP field. With a focus on areas like natural language understanding, information retrieval, and speech interfaces, these patents have evidently driven NLP forward, providing robust foundations for further innovation. This citation-based ranking is an essential aspect of patent analysis as it helps identify key patents that have significantly influenced the field.

In terms of the IPCR Classification that has received the most citations in the field of NLP, 'G06F17/30', which concerns 'Information retrieval; Database structures', ranks first. This classification has amassed a total of 606,597 citations, indicating that patents under this classification have had a significant impact in the NLP domain.

On the other hand, we can observe the total number of citations by year from Figure 8. Starting from the early 2000s, there is a noticeable increase in the number of patent citations, peaking in around 2007–2008. This rise could correspond to a period of significant innovation and development in NLP, leading to a high number of citations as these new ideas were referenced by subsequent patents.

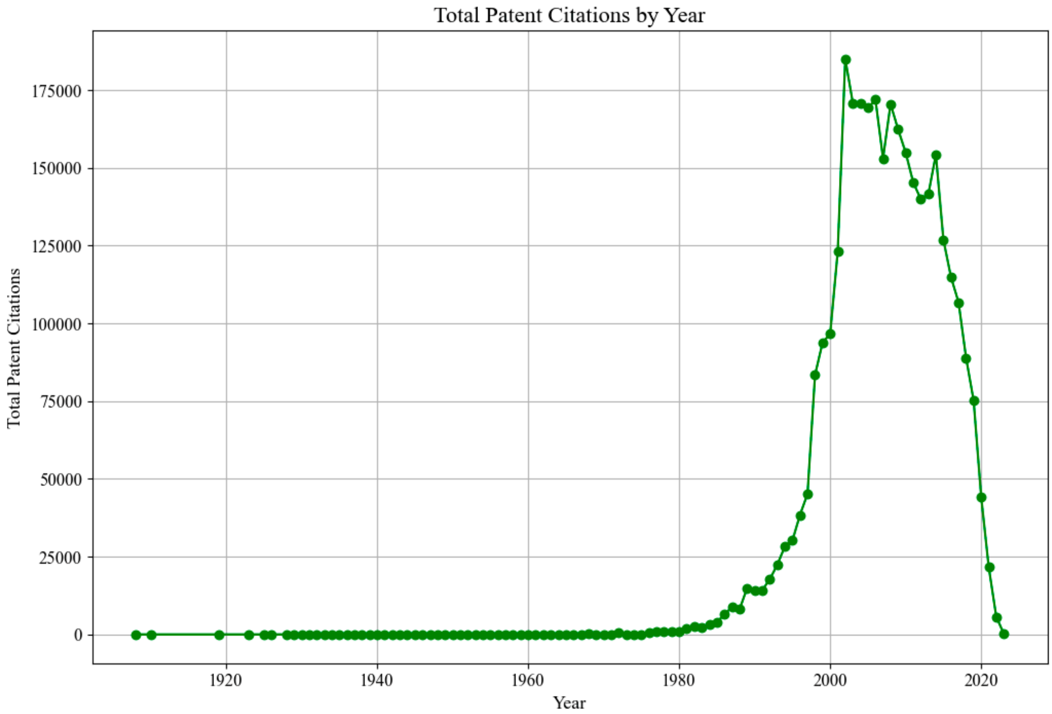

Figure 8. Total number of patent citations by year in NLP.

Post-2008, we observe a sharp decline in the total number of citations, which could be attributed to the relative novelty of the patents published during this period. These recent patents might not have had enough time to accumulate a significant number of citations.

4.6. NLP Abstract Shift over Time

In order to identify the trends and shifts in focus in the NLP patent landscape over time, we extracted the patents' abstracts and used topic modeling to understand the change of patent focus as follows:

1. Foundational Technologies (1990s): In the early 1990s, the focus was largely on foundational technologies for voice recognition and natural language processing. This is reflected in the prominence of words related to basic components of these systems, like 'voice', 'recognition', 'input', 'signal', 'dictionary', 'phoneme', 'sentence', 'word', 'knowledge', 'analysis', 'structure', 'language', and 'pattern'.

2. Internet Era (2000s): As we move into the 2000s, we see a shift towards the internet and networked systems. This is indicated by the emergence of words related to these themes, like 'client', 'server', 'service', 'network', 'communication', 'database', 'query', 'web', 'document', and 'search'.

3. User Interaction (2010s): In the 2010s, there is an increased emphasis on user interaction and multimedia content. This is seen in the prominence of words like 'user', 'interface', 'display', 'image', 'audio', 'message', 'agent', 'items', and 'resource'.

4. Machine Learning and AI (2020s): In recent years, we see a clear trend towards machine learning and AI, with a continued emphasis on user interaction technologies and connected systems. This is reflected in the prominence of words like 'model', 'feature', 'training', 'network', 'sequence', 'voice', 'recognition', 'signal', 'module', 'vehicle', 'control', 'device', 'input', 'display', 'audio', 'image', 'user', 'communication', 'content', 'module', 'vehicle', 'control', and so on.

These trends indicate the advancement in technology and the increased complexity of NLP systems over the years. From basic components of voice recognition and language processing systems in the early 1990s, the focus has shifted towards more complex and advanced themes related to machine learning, AI, user interaction technologies, and connected systems in recent years.

4.7. NLP Patents from Application to Publication

The time it takes for a patent to go from application to publication varies widely across different jurisdictions. On the faster end of the spectrum, we have South Korea, where a patent has been published just 7 days after its application. Similarly, in Spain, Denmark, Australia, and Germany, the shortest times from application to publication are 8 days, 12 days, 14 days, and 20 days, respectively. This suggests that these countries have efficient patent processing systems, which can be advantageous for inventors seeking quick protection and commercialization of their inventions.

However, on the other end of the spectrum, there are jurisdictions where the process can take much longer. In the United States, one patent took as long as 8799 days (about 24.11 years) to go from application to publication. Similarly, the European Patent Office, China, Germany, and Canada have seen their longest application to publication times reach 6274 days (about 17.18 years), 5755 days (about 15.77 years), 5434 days (about 14.88 years), and 5207 days (about 14.26 years), respectively. These extended durations could be due to various factors, such as the complexity of the patent, patent office backlog, patent oppositions, or extensive examination procedures. Despite the longer processing times, these jurisdictions are known for their rigorous patent examination processes, leading to more robust and enforceable patents. Overall, the average time from application to publication for the patents in the provided dataset is approximately 558 days.

Figure 9 illustrates the time that a patent takes from application to publication based on the IPCS class. The IPCR classification 'G06F40/00', which pertains to 'Handling natural language data', has the longest average time from application to publication, taking approximately 2.36 years.

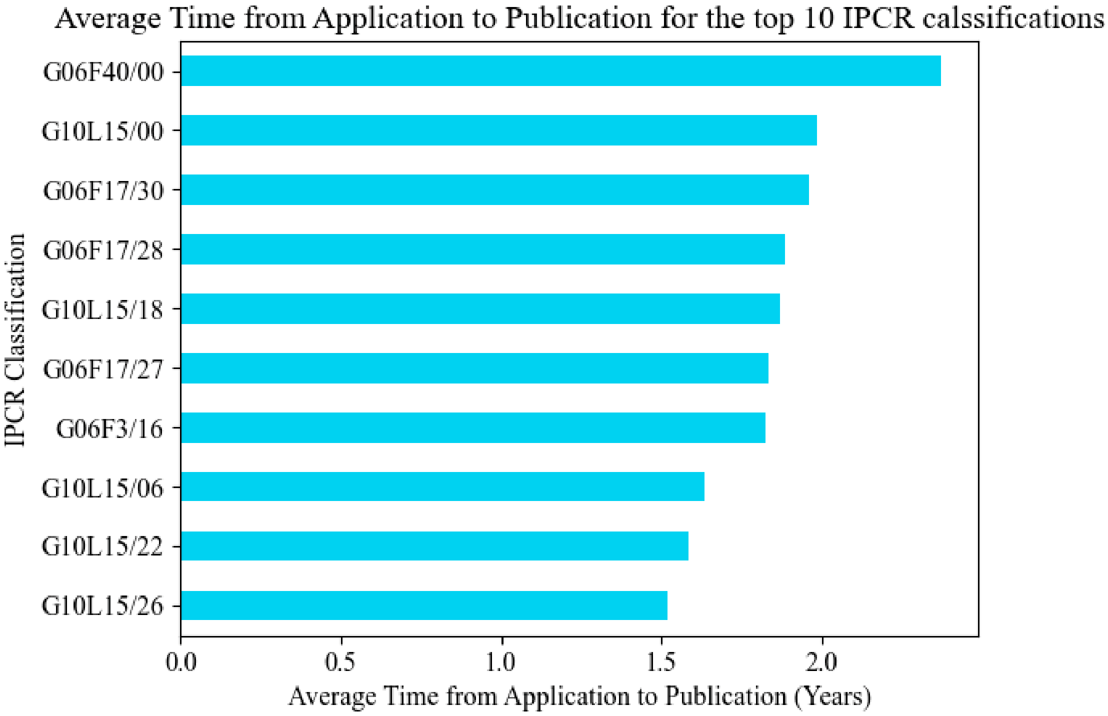

Figure 9. Average time from application to publication for patents within the top 10 most common IPCR Classification.

On the other hand, 'G10L15/26', associated with 'Speech to text systems', has the shortest average time at approximately 1.52 years.

The remaining IPCR classifications, including 'G10L15/22' (Speech recognition techniques), 'G10L15/00' (Speech recognition)'G06F17/30' (Information retrieval; Database structures), and 'G06F17/28' (Query processing), all exhibit varying average application to publication times, ranging between 1.5 and 2 years.

This figure provides a clear visual representation of the average time taken from application to publication for patents within these common IPCR classifications. It highlights the variability in this timeframe across different areas of focus within the field.

4.8. NLP Patents' Legal Status

Lens has categorized the legal status of patents into seven distinct categories, as depicted in Figure 10. The majority in the dataset were active, meaning they have been granted and are currently in force, according to Lens' definition. The bar chart shows that approximately 22% of the patents were in pending status, which is an indication of the application being under review. The expired status ranks third, which is given to patents that have reached the term date and are no longer in effect. On the other hand, discontinued patents refer to those that have been withdrawn, rejected, or discontinued for various reasons, potentially including payment-related issues. Inactive patents are granted patents that are not currently in force but have the possibility of reactivation, as they have not yet reached their term date yet. Their deactivation could be due to various reasons, which may include payment matters. Patented refers to patents which are registered and granted for protection internationally, namely by the World Intellectual Property Organization (WIPO). Unknown simply refers to not having enough information about the status. Definitions and more details for these statuses can be found on lens search filters[1].

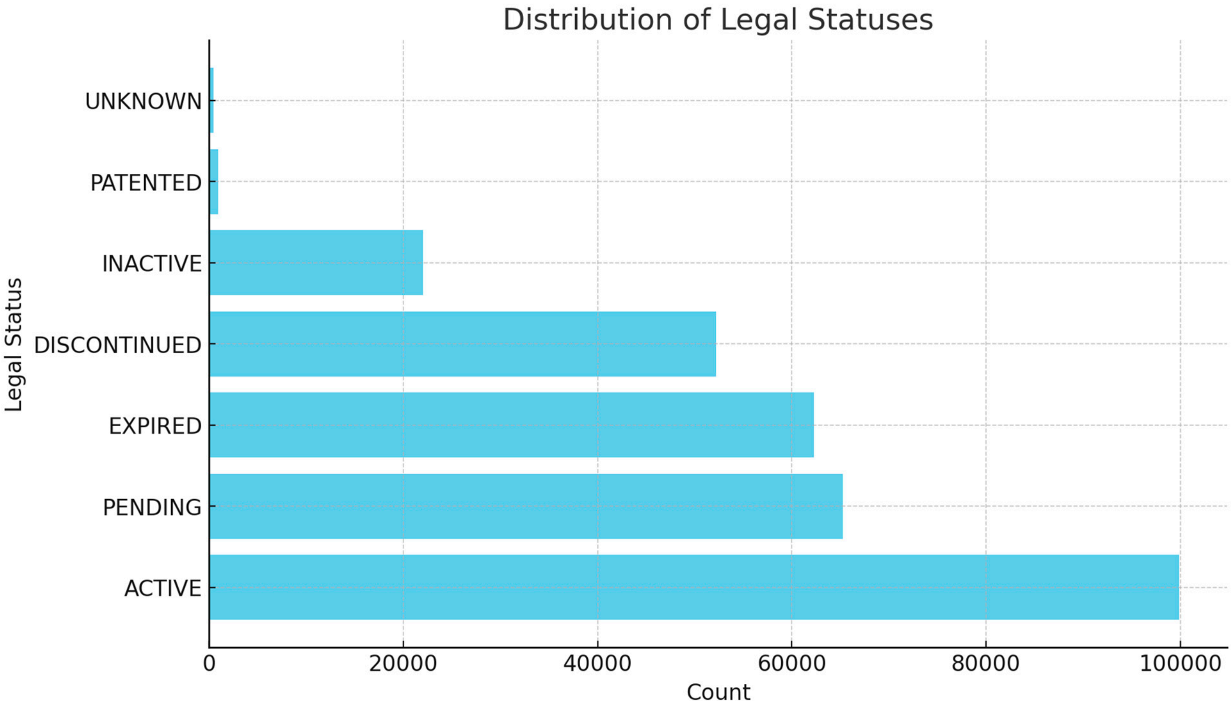

Figure 10. Distribution of legal statuses for the NLP patents in the dataset.

Likewise, examining Figure 11 provides insights into the evolving legal status of patents over the course of time.

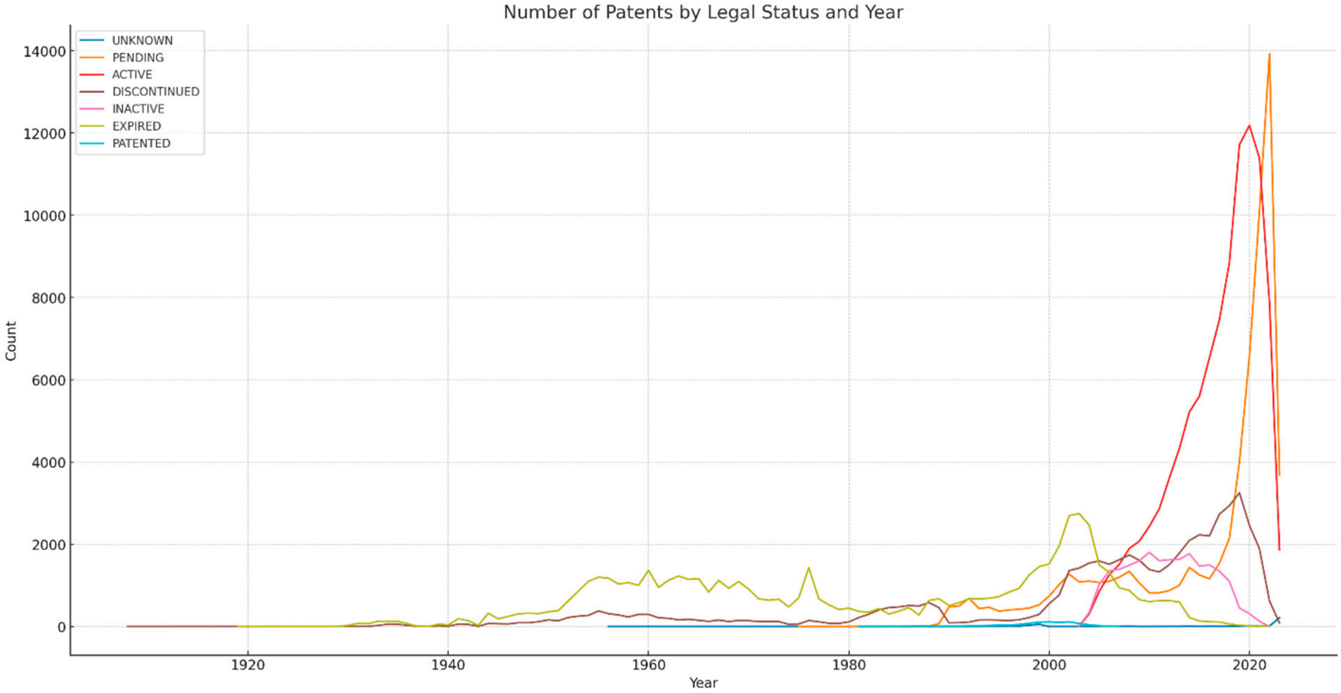

Figure 11. Number of NLP patents by legal status over the years.

The figure shows that the number of patents in the 'Active' status appears to be increasing over the years, indicating an increasing number of patents being approved and issued. Likewise, the number of 'Pending' status patents also seems to be increasing, but at a slower rate compared to the 'Active' status. This suggests a growing volume of patent activity, reflecting perhaps advancements in technology, increasing investment in research and development, or changes in patent filing behavior. In contrast, the 'Pending' status line consistently remains below the 'Active' status line from 2005 to 2020, indicating a slower growth rate and implying an increasing backlog of patent applications awaiting approval. This could be due to a variety of reasons such as increasing complexity of patents, limited resources at patent offices, or changes in patent examination procedures.

The number of patents in the 'Discontinued' status is relatively stable, with a slight increase in more recent years. This might be related to the overall increase in the number of patents, as a certain percentage of patents typically expire due to non-payment of maintenance fees. The 'Expired' status shows a sharp increase in one particular year, which might be due to a large batch of patents reaching the end of their maximum term in that year (typically 20 years). This reflects the natural life cycle of patents.

5. Conclusions and Outlook

In this paper we analyzed the global landscape of Natural Language Processing (NLP) patents and reported on its main technical approaches and trends. The recent increase in NLP patents, especially over the period 2018–2022 as previously observed, can be attributed to technological advancements in machine learning and AI that have opened new possibilities for NLP applications across a wide range of industries, leading companies to invest heavily in NLP research and patent filings to gain a competitive advantage and protect their intellectual property. Additional factors fueling growth in NLP patents include the greater availability of data to power NLP systems, and heightened research efforts and funding towards NLP across industry and academia—all driving more NLP inventions and patent filings, though the increase in patents may outpace the rate of actual innovation.

Furthermore, most of the patents related to NLP are filed in first world countries, where the US and China are leading the invention market. This could be due to the high-profile companies, such as IBM and Microsoft, that are based in these countries and have a major share of the patenting market.

As for the patent applicants, they can be divided into two main groups: industrial companies, and academic institutions. Out of the two, the industry sector dominates the NLP patenting market, with most applications belonging to Silicon Valley's technology giants that possess considerable resources and a strong commercial motivation to engage in research and development, obtaining patents in AI and NLP to safeguard their investments and maintain a competitive advantage. On the other hand, most academic institutions lack the funding, computational resources and commercial incentive that most companies in the industry sector have in abundance. This has led to a reduction in the number of NLP patents filed by academic institutions worldwide.

Moreover, this study also highlighted another notable contrast between the top industry entities and leading academic institutions in NLP patent volumes. In the industry sector, the prominent patent applicants are IBM, Microsoft, and Google, which are based in the United States, highlighting the significant influence of American technology companies on global innovation. Conversely, the academic sector is dominated by Chinese institutions, exemplified by the Institute of Computing Tech CN Academy. This signifies China's substantial investment in academic research and technology, showcasing the country's strategic commitment to enhancing its domestic technological capabilities.

Regarding the NLP topics or sub-domains that are most applied for and researched, this study found that the field of speech recognition is gaining an increasing interest in research and invention activities, which is supported by the findings of Chao et al. in [8]. The growing enthusiasm for speech recognition could be attributed to the recent commercial hype about personal assistants, such as Siri, and other conversational AI products.

Based on the analytical study we have conducted on the collected dataset of NLP patents, the following are the key findings of this paper, which could be used as a guide to assist researchers and inventors in applying for a patent in this field:

1. Timing: The average time from application to publication is around 500 to 600 days, but this can be shorter or longer. Knowing this information can help inventors in planning for their patent filing and application process beforehand and in being prepared for potential delays.
2. Jurisdiction: The United States and China are the top jurisdictions for NLP patents in this dataset. These jurisdictions could be important to consider for inventors considering international patent protection.
3. Collaboration: Collaborations appear to be common in NLP patents, as evidenced by the high number of patents with multiple inventors or assignees. Collaborating with others could be a beneficial strategy for patent applicants to consider, as it allows for pooling of resources and expertise.
4. Maintenance: A small portion of the patents in the dataset are in the 'Discontinued' status, indicating that they expired due to failure to pay maintenance fees. Inventors and patent applicants should be mindful of the ongoing costs of maintaining a patent once granted.
5. Trends: The number of NLP patent applications has been increasing over the years. This suggests that the field is growing and evolving, but also that competition might be increasing. Staying up-to-date with the latest research and developments in NLP can help identify unique and patentable ideas.
6. Legal Status: The majority of patents are granted, but a significant number are still in the 'Pending' status. This could indicate a backlog in the patent examination process, or it could suggest that some patents face challenges in meeting the requirements for grants. Inventors should consider thorough preparation of their application and potentially seek professional advice to improve their chances of a successful outcome.

The analysis conducted here provides a broad overview of the patent landscape in the field of NLP based on the data provided. However, there are several limitations that should be kept in mind:

1. Data Completeness and Accuracy: The analysis is limited by the completeness and accuracy of the data provided. For instance, if certain patents, inventors, or assignees are not included in the dataset, or if there were errors in the data, this would impact the analysis results.
2. Temporal Limitations: The dataset includes patents up to a certain point in time (the data cutoff). Therefore, trends observed in the data might not reflect more recent developments in the NLP field.
3. Generalization: The insights derived from this analysis are specific to the NLP field and the patents included in the dataset. They may not necessarily generalize to other technological fields or to all NLP patents.
4. Incomplete Technological Landscape: The dataset might not cover all aspects of NLP innovation, especially if some innovations are not patented or if they are protected through alternative means.

In conclusion, this analysis contributes to a better understanding of the NLP patent landscape, provides a methodological reference for patent data analysis, and offers valuable data and insights for further research and practice.

Author Contributions: Methodology, H.S.A.-K., T.A. and G.A.; Software, T.A.; Formal analysis, H.S.A.-K.; Data curation, T.A. and G.A.; Writing—original draft, H.S.A.-K.; Visualization, T.A. and G.A.; Supervision, H.S.A.-K.; Funding acquisition, H.S.A.-K. All authors have read and agreed to the published version of the manuscript.

Funding: This paper received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is available: https://github.com/iwan-rg/NLP-Patents (accessed on 13 November 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Note

[1] https://www.lens.org/lens/search/patent/list?q=nlp (accessed on 22 July 2023)

References

1. Natural Language Processing-Global | Market Forecast, Statista. Available online: https://www.statista.com/outlook/tmo/artificial-intelligence/natural-language-processing/worldwide (accessed on 21 July 2023).
2. Haney, B. Patents for NLP Software: An Empirical Review. *IUP J. Knowl. Manag.* 2020. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3594515 (accessed on 21 July 2023).
3. Warin, T.; Duc, R.L.; Sanger, W. Mapping Innovations in Artificial Intelligence through Patents: A Social Data Science Perspective. In Proceedings of the 2017 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 14–16 December 2017; pp. 252–257. [CrossRef]
4. Yoo, Y.; Lim, D.; Kim, K. Patent Analysis Using Vector Space Model and Deep Learning Model: A Case of Artificial Intelligence Industry Technology. *Preprints* **2021**, 2021110208. [CrossRef]
5. Ngoc, N.T.B.; Ngoc, H.T. Patent Relating to Artificial Intelligence and Liability for Artificial Intelligence Application from the US Law Perspectives. *Vietnam. J. Leg. Sci.* **2022**, *7*, 59–72. [CrossRef]
6. Haney, B. AI Patents: A Data Driven Approach; Rochester, NY, USA, 28 January 2020. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3527154 (accessed on 21 July 2023).
7. Leusin, M.E.; Günther, J.; Jindra, B.; Moehrle, M.G. Patenting patterns in Artificial Intelligence: Identifying national and international breeding grounds. *World Pat. Inf.* **2020**, *62*, 101988. [CrossRef]
8. Chao, M.-H.; Trappey, A.J.C.; Wu, C.-T. Emerging Technologies of Natural Language-Enabled Chatbots: A Review and Trend Forecast Using Intelligent Ontology Extraction and Patent Analytics. *Complexity* **2021**, *2021*, 5511866. [CrossRef]
9. Arts, S.; Hou, J.; Gomez, J.C. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Res. Policy* **2021**, *50*, 104144. [CrossRef]

10. Balsmeier, B.; Assaf, M.; Chesebro, T.; Fierro, G.; Johnson, K.; Johnson, S.; Li, G.C.; Lück, S.; O'Reagan, D.; Yeh, B.; et al. Machine learning and natural language processing on the patent corpus: Data, tools, and new measures. *J. Econ. Manag. Strategy* **2018**, *27*, 535–553. [CrossRef]
11. The Lens-Free & Open Patent and Scholarly Search. Available online: https://www.lens.org/lens (accessed on 22 July 2023).
12. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.