Datasheet for Indian Patent Dataset

Aryan Singh

April 2024

1 DataSheet

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to provide comprehensive information about patents filed in India. The specific purposes or tasks could vary, but generally, it serves several potential purposes:

- **Research and Analysis**: Researchers and analysts may use the dataset to study trends in patent filings over time, analyze the distribution of patents across different fields of invention, and assess the involvement of various countries and entities in the Indian patent system.
- **Policy Making**: Government agencies or policymakers might use the dataset to evaluate the effectiveness of patent policies, identify areas for improvement, and make informed decisions about intellectual property rights.
- **Business Intelligence**: Businesses and entrepreneurs could leverage the dataset to gather competitive intelligence, identify emerging technologies or market trends, and assess the patent landscape in specific industries.
- **Legal and Regulatory Compliance**: Legal professionals and patent attorneys may use the dataset to conduct prior art searches, assess the novelty of inventions, and ensure compliance with patent laws and regulations.
- **Academic Research**: Academics and scholars may utilize the dataset for empirical studies on innovation, technology transfer, and the economic impact of patents on various sectors.

Overall, the dataset fills the gap by providing valuable insights into the patent landscape in India, which can be utilized across various domains for decision-making, research, and analysis.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created solely by me, utilizing data scraped from the Indian Patent Advanced Search System (https://iprsearch.ipindia.gov.in/publicsearch).

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The creation of the dataset was not funded by any external entity.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances within the dataset represent individual patent applications filed in India. Each instance corresponds to a specific patent application and includes various attributes such as application number, title, application date, inventor information, applicant information, and application status.

There is only one type of instance in the dataset, which are the patent applications themselves.

Additionally, it's worth noting that each year's data is stored as a separate CSV file.

How many instances are there in total (of each type, if appropriate)?

The total number of instances in the dataset varies for each year:

- For the year 2010, there are 35,834 instances of patent applications.
- For the year 2011, there are 39,820 instances of patent applications.
- For the year 2019, there are 50,034 instances of patent applications.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset for the three years provided (2010, 2011, and 2019) is complete, containing all the patent applications filed during those respective years. However, it represents only a subset of the larger set, which includes all patents filed in India from 1950 to 2024 (current data).

Since the dataset is complete for each individual year, it can be considered representative for research and analysis within those specific years. However, it may not be representative of the entire patent landscape in India over the entire period from 1950 to 2024.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?

In either case, please provide a description.

Each instance in the dataset consists of structured features representing various attributes of individual patent applications filed in India. These features include:

- Application Number
- Title
- Application Date
- Status
- Publication Number
- Publication Date (U/S 11A)
- Publication Type
- Application Filing Date
- Priority Number
- Priority Country
- Priority Date
- Field of Invention
- Classification (IPC)
- Inventor Name
- Inventor Address
- Inventor Country
- Inventor Nationality
- Applicant Name
- Applicant Address
- Applicant Country
- Applicant Nationality
- Application Type
- E-MAIL (As Per Record)
- ADDITIONAL-EMAIL (As Per Record)
- E-MAIL (UPDATED Online)
- REQUEST FOR EXAMINATION DATE
- FIRST EXAMINATION REPORT DATE
- Date of Certificate Issue
- POST GRANT JOURNAL DATE
- REPLY TO FER DATE
- PCT INTERNATIONAL APPLICATION NUMBER
- PCT INTERNATIONAL FILING DATE
- Application Status

These features provide comprehensive information about each patent application, enabling detailed analysis and study of the patent landscape in India.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, there is a feature named "Application Status" associated with each instance, which serves as a label or target for classification purposes. This feature indicates the current status of the patent application, which could include categories such as "Granted Application," "Reply Filed," "Application is amended examination," "Abandoned," "Deemed to be withdrawn," "FER issued," or any other possible status.

Is any information missing from individual instances? If so, please

provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No information is missing from individual instances in the dataset. Every available piece of information linked to each patent application has been scraped and included in the dataset. Therefore, the dataset has all available information for each patent application, and no data is intentionally removed or redacted.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

In the dataset of patent applications for India, there are no explicit relationships made between individual instances. Each patent application is treated as an independent entity, and there are no direct connections or links established between them within the dataset. Therefore, relationships between individual instances are not made explicit.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There are no recommended data splits provided for this dataset.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There are no errors, sources of noise, or redundancies present in the dataset.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained and does not rely on or link to external resources.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

Yes, the dataset contains individual personal information of those who have filed the patents, which may be considered confidential. This information could include names, addresses, email addresses, and other identifying details of inventors and applicants. As such, it's important to handle this data with appropriate care and consideration for privacy regulations.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No, the dataset does not contain data that, if viewed directly, might be offensive, insulting, threatening, or otherwise cause anxiety.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset contains information about people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No, the dataset does not identify any subpopulations such as age or gender.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Yes, it is possible to identify individuals directly from the dataset as their name, address, country, nationality, and email information are provided.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No, the dataset does not contain data that might be considered sensitive in any way.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data for each instance was scraped directly from the Indian Patent Advanced Search System (https://iprsearch.ipindia.gov.in/publicsearch). This data was directly observable from the search portal.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data was collected by automating the process of patent search and downloading of tables from the search portal using Selenium and Python.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

No sampling strategy was employed as the dataset represents complete data for the specified years.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g.,

how much were crowdworkers paid)?

No individuals were involved in the data collection process. The data was collected solely by me.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data was collected over the timeframe corresponding to the years 2010, 2011, and 2019, as specified in the dataset. This timeframe matches the creation timeframe of the data associated with the instances.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No ethical review processes were conducted for this dataset. However, ethical review is planned for the future once all the patent data for all the years is collected.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset does relate to people.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was collected directly from the Indian Patent Search Portal website. No direct contact with individuals was made.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Given the nature of the data collection process, which involved scraping information from a public website, it wasn't feasible to notify individuals about the data collection. Contacting potentially hundreds of thousands of individuals would not have been practical or possible.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

The individuals whose data is included in the dataset filed for patent applications, which is publically available information as per Indian laws, such as the Indian Patents Act, 1970. Hence, there is no explicit consent required for the collection and use of this data.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data pro-

tection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

While explicit consent was not obtained due to the nature of the data collection process, individuals who wish to revoke any potential consent or have concerns about the use of their data can contact us directly.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Some basic preprocessing was performed on the data, including cleaning the strings and removing unnecessary symbols. Otherwise, the data is ready to use without further preprocessing.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Since the data saved is nearly the same as raw, the raw data was not explicitly saved.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

The preprocessing was done using Python along with libraries such as Pandas and scikit-learn, all of which are publicly available.

Python: https://www.python.org/
Pandas: https://pandas.pydata.org/
scikit-learn: https://scikit-learn.org/stable/index.html

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

The dataset has been utilized to analyze the patent filing patterns of individuals and large firms, particularly focusing on the number of patents filed and the success rate of obtaining granted patents.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No, there is no repository that links to any papers or systems that use the dataset.

What (other) tasks could the dataset be used for?

The dataset could be used for various tasks related to patent analysis and intellectual property research, including:

1. Trend Analysis: Studying trends in patent filings over time across different technology fields.

2. Innovation Mapping: Identifying emerging technologies and areas of innovation based on patent filings.

3. Competitive Intelligence: Analyzing the patent landscape

to gain insights into competitors' activities and technological strengths.

4. Policy Evaluation: Assessing the effectiveness of patent policies and regulations in promoting innovation and economic growth.

5. Technology Transfer: Studying patterns of patent ownership and collaboration to understand technology transfer dynamics.

6. Patent Valuation: Using patent data to assess the value and potential commercialization opportunities of inventions.

7. Legal Research: Conducting prior art searches and assessing the novelty and patentability of inventions.

8. Economic Analysis: Investigating the economic impact of patents on industries and regions.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Future users should carefully review the license of the dataset before use and ensure compliance with laws regarding the handling of personal information of individuals. It is prohibited to sell or distribute this data. Additionally, users should avoid using the dataset in any manner that could result in unfair treatment of individuals or groups, such as stereotyping or quality of service issues. To mitigate these undesirable harms, users should consider implementing safeguards such as anonymization or aggregation of data where possible, and adhere to ethical guidelines and legal regulations governing the use of personal data.

Are there tasks for which the dataset should not be used? If so, please provide a description.

There are no specific tasks for which the dataset should not be used. However, users should ensure that their use of the dataset complies with laws regarding the handling of personal information and does not violate individuals' privacy rights.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset will be distributed to third parties; however, it will not be open source. Distribution will only occur after companies or academic institutions provide complete affidavits or proof that they will not misuse the dataset and will adhere to all licensing laws concerning privacy issues.

How will the dataset will be distributed (e.g., tarball on website,

API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset will be distributed exclusively through GitHub of Dataset after a thorough verification process, including submitting affidavits ensuring responsible use. Currently, the dataset does not have a Digital Object Identifier (DOI), but one will be obtained after data collection for all the years.

When will the dataset be distributed?

The dataset is currently available for distribution.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset will be distributed under a Creative Commons Attribution-ShareAlike (CC-BY-SA) license. This license allows for the use of the dataset only after verification has been completed and prohibits the sale or sharing of the data with anyone who has not been part of the verification process. Additionally, the terms of use (ToU) will include provisions to protect individuals' personal information from being leaked or misused. Copyright will be retained by the entity responsible for creating the dataset.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No third parties have imposed any IP-based or other restrictions on the data associated with the instances.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Since we're using a Creative Commons Attribution-ShareAlike (CC-BY-SA) license, there are generally no export controls or regulatory restrictions that apply to the dataset or individual instances. This license allows others to distribute, remix, adapt, and build upon the dataset for any purpose, even commercially, as long as they give appropriate credit, provide a link to the license, and indicate if changes were made. If others remix, adapt, or build upon the dataset, they must distribute their contributions under the same license as the original.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The owner will be responsible for supporting, hosting, and maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The owner, which is me, can be contacted via email at `aryansinghmain09@gmail.com`.

Is there an erratum? If so, please provide a link or other access point.

No, there is no erratum available for the dataset at this time.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes, the dataset will be updated periodically to correct labeling errors, add new instances, or delete instances as necessary. Updates will be communicated to users through GitHub, where the dataset is hosted. The frequency of updates will depend on the availability of new data and any changes required to improve the quality of the dataset. Updates will be managed by the dataset owner (myself) and will be made available to users as soon as they are implemented.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Since there are no applicable limits on the retention of the data associated with the instances, individuals were not informed of any fixed period of time for data retention or deletion. Therefore, no specific limits or enforcement measures are in place for data retention in this regard.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Older versions of the dataset may not continue to be actively supported, hosted, or maintained. However, they may still be accessible in the dataset repository for reference purposes. If older versions become obsolete and are no longer supported, this information will be communicated to users through the dataset repository, indicating that users should transition to newer versions for updated data and support.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

No, there is no mechanism for others to extend, augment, build on, or contribute to the dataset. Contributions from external parties are not accepted, and there is no process for validating, verifying, or distributing contributions to other users.