Journal of Scientometric Research, 2025; 14(3):889-908.
https://www.jscires.org

Research Article

# Measuring healthcare Innovations through Human Assigned Approach and Man-Machine Derived Approach: A Comparative Analysis Using Published Patents of India

**Debasis Majhi[1], Priya Tiwari[2], Bhaskar Mukherjee[3,*]**

[1]Kalna College, West Bengal, INDIA.
[2]Tagore Library, University of Lucknow, Lucknow, Uttar Pradesh, INDIA.
[3]Banaras Hindu University, Varanasi, Uttar Pradesh, INDIA.

## ABSTRACT

**Aim:** This study examines the effectiveness of the Latent Dirichlet Allocation (LDA) model in extracting thematic structures from healthcare patents and compares machine-generated topics with human-assigned International Patent Classification (IPC) codes. It also assesses whether using both patent titles and abstracts improves topic identification compared to titles alone. **Methodology:** Healthcare-related patents published in India between 2000 and 2022 were retrieved from the WIPO PATENTSCOPE database. IPC classifications served as the benchmark for human-assigned categorization. LDA-based topic modeling was applied to patent titles, abstracts, and their combined text, and the resulting topics were compared with IPC classifications to assess alignment and thematic coverage. **Findings:** IPC analysis identified key innovation areas, including medicinal preparations, organic active ingredients, and herbal drugs. LDA applied to titles highlighted themes such as crystalline pharmaceutical and herbal compositions, while abstract-based analysis revealed more detailed topics, including antiviral agents and rotavirus vaccine compositions. Although LDA effectively extracted latent topics, title-only analysis provided limited thematic depth. **Implications and Recommendations:** Combining patent titles and abstracts significantly improves the accuracy and comprehensiveness of LDA-based topic modeling. While machine learning supports large-scale patent analysis, human expertise remains crucial for interpreting results and refining trend analysis. Hybrid analytical approaches are therefore recommended. **Contribution and Value Added:** The study confirms the usefulness of machine learning for healthcare patent analysis in the Indian context. It adds methodological value by demonstrating the benefits of multi-field textual input and highlights the complementary roles of automated models and expert judgment in patent analytics.

**Keywords:** Healthcare patents-India, Latent Dirichlet Allocation, LDA, Machine extraction-patent topics, Patent Classification, Topic Modelling-Patents.

**Correspondence:**
**Bhaskar Mukherjee**
Department of Library and Information Science, Banaras Hindu University, Varanasi, Uttar Pradesh, INDIA.
Email: mukherjee.bhaskar@gmail.com
ORCID: 0000-0003-2077-6976

**Received:** 02-09-2025;
**Revised:** 14-10-2025;
**Accepted:** 24-11-2025.

## INTRODUCTION

Patents are crucial indicators of a country's innovation to understand how far a nation's technology has advanced and, perhaps more importantly, how those advances have affected development. The patent system also promotes innovation by sharing patent information for further advancements. It balances commercialization and access through exclusive rights and voluntary licensing, while incorporating flexible mechanisms like compulsory licenses and research exceptions to serve public interests.[1] Whenever an inventor develops a new technology, design, or process, he(s) has to apply to the government patent office of the nation for the issuance of patents. The national patent office maintains a database of the applied/ granted patents. Some of the largest and most reliable patent office databases include the Canadian Patents Database maintained by the Canadian Intellectual Property Organization (CIPO), DEPATISnet hosted by the German Patent and Trademark Office (DPMA), Espacenet produced by the European Patent Office (EPO), JP-PlatPat produced by the Japan Patent Office (JPO), U.S. Patent Assignment Database (USPTO), inPASS by the Indian Patent Design and Trademark Office. PATENTSCOPE, a worldwide patent database maintained by the World Intellectual Property Organization (WIPO), provides free online access to all international patent applications filed under the Patent Cooperation Treaty (PCT) as well as all relevant papers and patent collections from National and Regional Offices (WIPO Guide to Using Patent Information, 2021).[2]

**DOI:** 10.5530/jscires.20251617

The health sector is one of the most significant investors in innovation after the Information Technology (IT) sector.[3] In all nations, a sizeable share of annual private and public R&D spending is on health research and development. Prevention is continually being improved by innovation. For instance, the novel mRNA vaccines enable primary COVID-19 prevention, while the identification of circulating tumour cells[4] enables secondary COVID-19 prevention. Almost all new innovations and medical advancements are published or filed in the form of patents, which is a suitable yardstick for measuring medical innovation. Policymakers face the challenge of finding the right balance between patent owners' rights and public needs.[5] Developing new drugs requires significant investment, research,[6] and clinical trials, and patents provide an incentive for such endeavors.

According to the Global Innovative Index, 2023,[7] India ranks 40, slightly higher than its earlier rank of 46 in 2021 but a huge leap from its rank of 81 in 2018. Healthcare innovation is one of the vital realms for India, as our country suffers from the 70:70 paradox, which means 70% healthcare expenditure is borne by the people out of their own pockets, of which 70% is spent on medicine alone.[8] According to the World Bank Report-2020, the public healthcare expenditure in India stands at only 2.96% of the gross domestic product, against the figures of 10.90% in Japan and 5.59% in China. The huge gap between the accessibility and affordability of healthcare services provides a huge opportunity for cost-effective, reliable, and affordable innovations for the betterment of society. The 'Make in India' initiative has now shifted to building products locally from scratch, rather than involving foreign companies to sell their products in India. Along with multinational companies, many local and regional companies are emerging in the Indian healthcare sector by establishing manufacturing units and incubation centres, which can be attributed to India's success story in pharmaceuticals industry in the past. At this point of juncture, it is important, therefore, to evaluate where India's healthcare innovation stands.

The dramatic increase in patent filing over the past two decades has necessitated a standardized system for classifying and organizing patents according to their technical content. The International Patent Classification (IPC) plays a significant role in identifying patents by assigning unique codes to patents based on their subject matter. These codes, consisting of almost 70,000 alphanumeric symbols, are based on a hierarchical structure of sections (A to H), classes (A61), subclasses (A61K), groups (A61K 6/00), and subgroups (A61K 6/52) of an area of invention. The subgroup title defines a field of subject matter within the scope of its main group (Guide to the International Patent Classification (2023). It ensures that patents from different countries and regions can be easily categorized and searched based on their subject matter. National property offices that do not have sufficient expertise for classifying to a detailed level are allowed to classify patents only in the main group. However, when a single invention is characterized by two or more inventive aspects so that each are depending on the contents of the patent document, more than one classification symbol may be allotted to a single patent. For e.g., a patent related to*Skin Lightening Composition* may be classified under A61K 8/36; A61K 8/97; A61Q 19/02; A61K 8/362.

Assessing innovative trends through patent-related information for a country on a specific subject can be possible by accessing international databases like USPTO, EPO, InPASS or PATENTSCOPE. Among these databases, except inPASS or PATENTSCOPE, no other databases have an exhaustive coverage of Indian patents. Despite inPASS being an authentic source for patent-related information of India, they do not allow downloading data as a chunk for understanding trends. In contrast, PATENTSCOPE of WIPO allows downloading data for a maximum of 10,000 records for its registered users. The metadata of downloaded records contains important fields like Application date, Country of Filing, Inventor name with address, Title, Abstract, IPC derived classes of the patents, but not keywords. Observing the existing literature corpus, we found that Chae and Gim (2019) proposed a model for analyzing the technical innovation of applicants based on International Patent Classification (IPC) and the Cooperative Patent Classification (CPC) classification scheme.[9] This study showed that these classification schemes can be used to extract the trends in applicants' technological innovation and to track changes in the innovation patterns. Although IPC classes of patents give fair ideas of areas of innovation, one has to possess expert knowledge to understand the IPC classes to perform the trend analysis. Additionally, an increasing number of patents contain multiple sub-group class from which identifying the core concept is another obstacle in processing and visualizing patent data through IPC classes.

This is exactly where machine learning algorithms like text mining and natural language processing come into play. Keeping in mind the abundance of unstructured data and language fragments, unsupervised approaches like topic modelling are employed to successfully extract topic keywords from a document through a semantic generalization of the document's content. Topic modelling algorithms are instrumental in extracting latent topics from published patents, offering valuable insights and enhancing the analysis of patent data. By employing techniques such as Latent Dirichlet Allocation (LDA), topic modelling algorithms can identify and uncover the underlying themes or subject areas within a large collection of patents. These algorithms analyze the textual content of patents, such as titles, abstracts, claims, and descriptions, and identify recurring patterns of words and phrases. By assigning probabilities to each word's association with specific topics, topic modelling algorithms can generate a comprehensive representation of the latent topics present in the patent collection.[10] These algorithms also facilitate the discovery

of emerging or niche topics in patent datasets. We therefore propose an alternative text-mining approach based on an unsupervised machine learning algorithm using Latent Dirichlet Allocation (LDA).[11] LDA is a type of unsupervised machine learning algorithm used to identify topics present in a corpus of text. It is a generative model that assumes each document is a mixture of topics and that each word's presence is attributable to one of the document's topics.[12] By detecting latent themes, LDA algorithms help uncover trends, identify areas of interest, and support decision-making processes for researchers, analysts, or organizations.[13] Gensim, a Python library that eliminates ambiguous phrases, helps build the LDA model. It is widely used in text mining, natural language processing, and other related fields. LDA works by assigning each document a set of topics based on the words that appear in it. In the recent times, LDA has been applied to the US healthcare data to identify phenotypic topics.[14] A study has also been conducted to identify latent topics from a set of documents, analyze long-term topical trends, and jointly model words and references in a document.[15]

The healthcare domain generates vast amounts of complex and unstructured data, including medical records, research articles, clinical notes, and patient feedback.[16] LDA topic modelling provides a powerful tool to extract meaningful insights from this data by identifying latent topics and themes.[17] It enables researchers to uncover hidden patterns, understand the relationships between medical concepts, and gain a comprehensive view of healthcare-related information. LDA topic modelling facilitates knowledge discovery, trend analysis, and the identification of emerging topics in healthcare research. This can aid in improving patient care,[18] clinical decision-making,[19] and disease management.[20] Application of LDA approach for patent analysis has been discussed to explore the development status of patent of ship integrated power system in China,[21] to identify relationship between terms and topics in industrial Cyber-physical system,[22] to identify technological topics in smart manufacturing,[23] trend analysis on green-house gases,[24] automatic patent classification system,[25] 5G telecommunications,[26] and humanoid robot[27] etc.

Although healthcare innovation has been widely studied around the world, most researchers still tend to rely on either expert opinions or automated patent-analysis tools, with very few trying to understand how these two approaches compare when used on the same set of data. In India, where healthcare innovation is growing rapidly, this comparison becomes even more important. Yet, the existing literature shows a clear gap: there is no empirical study that evaluates Indian healthcare patents using both human-assigned and machine-generated outputs in a systematic way. Many earlier studies look at broad technological trends or depend heavily on automated techniques, but they seldom check whether these machine-generated insights align with the understanding of domain experts. This can lead to inconsistencies or misinterpretations. To address this gap, the present study adopts a dual-method approach to examine healthcare-related patents published in India. The novelty of the study lies in bringing together the strengths of human expertise and computational analysis to see how each method captures the features of innovation. By comparing them directly, this study aims to build a more reliable and well-rounded way of measuring healthcare innovation.

Our approach differs substantially from the approaches used in earlier research. We first identified the corpus of patents from WIPO database by using the IPC classes related to healthcare. The class number of a patent is assigned by the patent examiner, generally an expert in the field, by following classification rules. On reading the patent document in detail, the classifier assigns a single or multiple sub-group number, which represents the thought content of the patent. From this corpus, major fields of innovation for a country like India in the healthcare domain have been identified. Further, how far university-industry-government are participating in the innovation process has also been explored. The same raw corpus was then used in LDA model and trained the machine to compute the number of topics regarded as the best representatives of the corpus correctly. On the basis of machine assigned number, topics were chosen from the title and abstract fields, and an umbrella topic was given based on the dominant terms in each topic. Finally, the computer-assigned title and abstract topics were compared first and then with the manually-assigned IPC-based top areas. While inter-field similarity, i.e., topics from titles with topics from abstracts, assists us in arriving at the conclusion whether title and abstract fields are complementary to each other in deriving latent trained topics of a subject, and how exhaustive the LDA-based topics are to track the trend in research in a subfield like healthcare.

Guided by the aim of comparing human-assigned and machine-derived approaches for assessing healthcare innovations in India, this study seeks to explore how these two methods differ in identifying and interpreting innovation within published healthcare patents. It examines whether their evaluations tend to align or diverge when applied to the same dataset, and what these similarities or differences reveal about the strengths and limitations of each method. Through this comparison, the study also aims to understand how reliable each approach is on its own, and whether combining them can offer deeper and more balanced insights into innovation measurement in the Indian healthcare sector.

## OBJECTIVES

Based on the identified gaps, the study pursues the following objectives:

- To identify major domains of healthcare innovation in India using expert-assigned IPC classes.

- To extract machine-derived latent topics from Indian healthcare patents using LDA applied to titles and abstracts.
- To compare machine-generated topics with human-assigned IPC classifications to evaluate alignment, divergence, and thematic accuracy.
- To examine the participation of universities, industry, and government sectors in healthcare patenting in India.
- To assess whether machine-derived topics provide additional insights beyond traditional human classification systems.

## METHODOLOGY

### Data and Source

As the present research was designed to identify the major areas and players in medical-related innovation, first, we consulted the International Patents Classification (IPC) to identify classes related to this domain. The IPC classes namely A61 (medical or veterinary science; hygiene) and its sub-class A61B (diagnosis; surgery; identification), A61J (medical or pharmaceutical devices or methods for bringing pharmaceutical products); K (preparations for medical, dental or toiletry purposes); L (methods or apparatus for sterilising materials, disinfection surgical articles); M (devices for introducing media into, or onto, the body, devices for producing or ending sleep or stupor); A62B (devices, apparatus or methods for life-saving); A01L (preparation of organic disinfectants for medical, dental, toiletry use); B01L (medical diagnostic apparatus), and G16H (healthcare informatics) have been chosen. A multiple search string consisting of IPC class, Applicant Address (IN) and Publication Time (01.01.2000 TO 01.01.2023) was fetched in the Patent Search database of World Intellectual Property Organization (WIPO). Despite, Indian Patent Advanced Search System (inPASS) is an alternative option for searching Indian Patents, the erratic results of inPASS further restricted us from using the WIPO database. The search was made during the first quarter of 2023.

All the results from individual IPC classes were downloaded, merged, and then duplicates were removed to identify unique patent titles in the above-mentioned IPC classes. A total of 10804 unique patents were finally selected for analysis that were published from January 2000 to December 2022. The applicant details, IPC classes, and Year of publication of all the unique records were imported to a Python-based program to know the major IPC classes and players, their connection with each other, and the areas of innovation they made. It was seen that a number of patent titles contained multiple classes. In such cases, the first two subgroup classes have been considered, and for counting total patents under a subgroup, equal weightage has been given to both subgroups. In this study, the major inventors have been grouped in the form of a triple helix structure, i.e., university-industry-government, rather than individual persons. When a patent was published by more than one of these three sectors, weightage has been given equally to both sectors for a single patent. To establish a relation between two IPC classes, Gephi 0.10.0 was used. The reason behind choosing Gephi was its flexibility to import data in different formats.

### Pre-processing for topic modelling

For topic modelling, first, the title and abstract fields of raw unique records were pre-processed. In this stage, we started pre-processing of the exported dataset to remove the unwanted, unnecessary information that was included in the dataset. We used the *Natural Language Toolkit* (NLTK) library for pre-processing the dataset. *NLTK* helped to remove undesirable words, characters, and symbols from the corpus and enhanced the dataset quality for analysis. In this step, we first converted all text into lowercase and removed all numbers, double spaces, and special characters. Then, we ran the correct function from the *TextBlob library* to perform spelling correction. Then, the *stopwords* that do not help to add much information to the text like 'a', 'an', 'the', 'but', 'of', and 'in' were removed from the text corpus using *NLTK* English language. After that, *nltk tokenization* was performed with the text corpus, considering each word as a single token. After that, the lemmatization procedure was performed with the help of WordNetLemmatizer for context analysis, which helps to reduce the inflectional form of words, thus avoiding different forms of word features. The *lemmatized* data were used to generate a dictionary, and the documents were vectorized using a *bag-of-words* algorithm that determined the frequency of terms in each document. The dictionary was generated using *Gensim's corpora dictionary*, and the corpus was built using the *dictionary.doc2bow's* integer encoding. It is important to mention that we have also employed stemming along with lemmatization for the purpose of reducing inflected forms of words to a common base form.[28] Because of the fact that stemming mostly collapses derivationally related words while lemmatization commonly only collapses the different inflectional form of lemma,[28] we preferred lemmatization only. In total, we get 6480 unique tokens in the title field and 21064 in abstract field. After the pre-processing of the dataset, each token was assigned a weight, and these data were sent to the LDA model. Figures 1a and b show the distribution of a number of tokens in the patent after pre-processing.

### Preparation of the topic model

In the next stage, we used probabilistic topic modelling, Latent Dirichlet Allocation (*LDA*) for identifying predictive and latent topics of our dataset. The *LDA* model was first trained using various model hyperparameters like chunksize, passes, alpha, beta etc. for the different k-number of topics, and finally, that number of topics was chosen that showed the lowest coherence score. To calculate the coherence score, *u_mass* coherence techniques were adopted, and the *matplotlib* library was used to

plot the graph. Figures 2a and b represent the coherence graph. As the coherence score seems to be decreasing with the number of topics, it makes better sense to pick the model that gave the least u_mass score. Here, the least coherence score is shown as 22 for titles and 23 for abstracts. If the number of topics selected is too large, it may lead to greater subdivision of topics and higher similarity between topics. On the other hand, if the number of topics is too few, it may lead to a number of merged topics. Since the focus is more on comparing machine-generated topics under various meta-tags, we, in the present study, have chosen 20 topics each from the title and abstract with the highest coherence score to compare the latent topics of our dataset.

Important to note that despite spending 350 min in a PC with an Intel Corei5 4590 CPU and 32 GB RAM we could not get any coherence score through c_v. In every instance, the code was halted. One possible reason was the data sparsity for a single word like 'Granules', 'inhibitors', 'Micropump' etc., as titles. As a result, our code was less susceptible to the data sparsity issue. Because of that, we had adopted u_mass for the coherence score. The Coherence Score for the title corpus and the abstract corpus were -15.49283 and -15.574885, respectively. Finally, to visualize the topic with the relative occurrence of each term therein, pyLDAvis library was used.

After generating the top twenty topics for each (title and abstract) category using the LDA topic modelling method, a comparative analysis was made with the manually top twenty IPC subjects in Table. LDA topic modelling is an unsupervised machine learning technique that reveals the hidden topic from the dataset. Thereby, LDA topic modelling techniques help to compare the man-made patent classification to machine-generated patent topics.

## RESULTS

This part of the results discusses the top domains of IPC classes under which almost 40% patents have been published, interconnections between these IPC classes, and top ten university-industry or government in India that played a leading role in producing these patents. As shown in Table 1, A61K or A61P i.e., devices or methods specially adapted for bringing pharmaceutical products, are two major areas of innovation in India. Under this area, preparation of lozenges, dragees, coated pills-based medicine/drugs, including herbal medicine for various disorders, are major form of innovation. Drugs, chemically made from heterocyclic compounds and used for disorders of metabolism, hyperglycemia, the nervous system, or cancer, were assumed to be the major domains of discovery during the last two decades. This is followed by the preparation of various cosmetic items for hair, skin or toiletry purposes.

While looking at the interconnections between the IPC classes in Figure 3, it has become clear that most of innovation under the class A61K 9/20 i.e. medicinal preparation, was in form of pills, lozenges etc. or powders (A61K 9/14) or agglomerates; granulates; microbeadlets (A61K 9/16) or dragees; coated pills or tablets (A61K 9/28) or capsules (capsules 9/48) as the lines between these nodes are thicker than other nodes. Similarly,

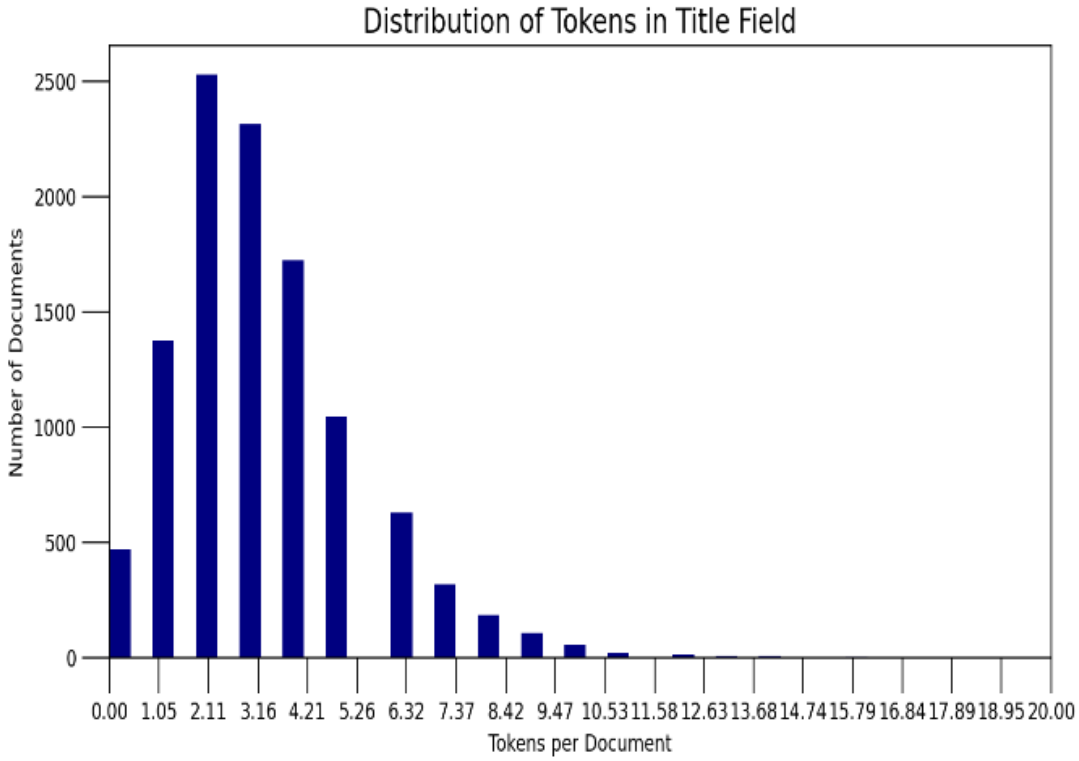


**Figure 1a:** Distribution of Tokens in Title Field.

the Antineoplastic drugs (A61P 35/00), i.e., drugs for cancer treatment, have a strong relation with patents bearing class number for compounds having heterocyclic rings consisting of nitrogen, oxygen, sulphur, selenium etc., and with no ring with steroids, saccharides, or peptides. On the other hand, the patents group A61Q - Cosmetics or Similar Toiletry Preparations have linkage with 39 other patent groups, mainly derivatives from algae, fungi, lichens, plants or obtained by reactions other than carbon-to-carbon unsaturated bonds, or containing vitamins or carboxylic acids, salts, or anhydrides.

While analyzing these patents in relation to their inventors, it was seen that commercial industries are playing a major role in innovation. Companies handling the manufacture of cosmetic goods for human necessity or pharmaceutical companies preparing drugs are the main industrial players in the healthcare sector. Most of these multinational companies are not only showing significant growth in the international market, but are also investing resources towards the development of goods as per the needs of the local market.

On the other hand, although quite less in number, Indian Institute of Science, Indian Institute of Technology from the academic sectors are some institutions who are associated with healthcare innovation as mentioned in Table 2. And, Council of Scientific and Industrial Research, Department of Biotechnology are among the major government setups involved in discovering healthcare issues.

## Latent topics analysis

This part of the analysis is based on machine learning techniques to analyze a large set of titles and abstracts data of patents and discover the set of topics in them. While handling our dataset, we observed that the existing stopword list of NLTK and spacy's "en_core_web_sm" is insufficient to exclude the unnecessary words for a dataset like ours. Therefore, it was decided to develop a list of domain-specific stopwords and then append them to the existing list. The list of a few such stopwords is available in Annexure I.

Using Gensim package for topic modelling in our dataset, we gathered 20 topics both from the titles and abstracts of patents without significant overlap. The list of those topics, along with the percentage of tokens on each term of a topic, is mentioned in Annexure II. As a topic is a collection of dominant keywords (tokens) that are typically representative of the whole and follow Dirichlet probability distribution, an umbrella topic has been assigned based on each topic's dominant keywords using the author's expertise. To confirm our understanding of the machine extraction, we set λ value of a topic in pyLDAvis visualization graph between 0.8 and 0.6 and noted only those most representative tokens that gained a higher score among others. Then these tokens were joined under one umbrella name.

The topic visualization graph of our dataset using pyLDAvis is shown in Figures 4a and b. In pyLDAvis, each bubble represents a topic. The larger the bubble, the higher the number of patent titles in the corpus, which indicates popularity. Also, the size of the bubbles shows the more prominent topics. Figure 4a shows that topic ID 1 (#15-Annexure II) is the most prevalent topic

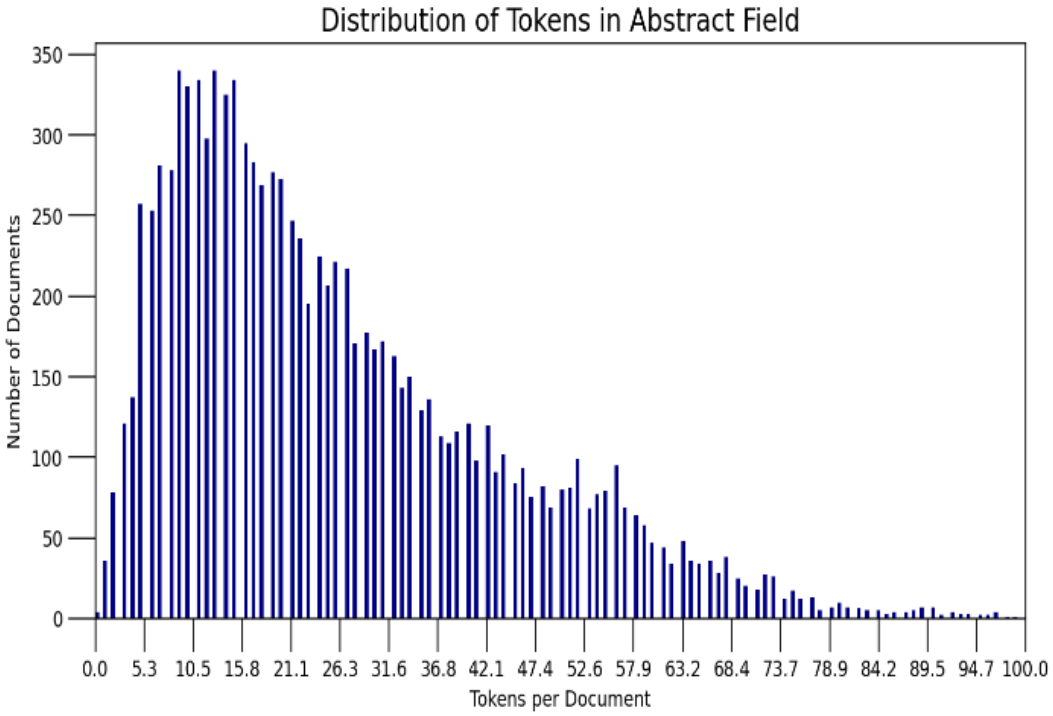


**Figure 1b:** Distribution of Tokens in Abstract Field.

in the title area, which indicates that the preparation process of anti-inflammatory pharmaceutical compositions/ salts, and from the abstract area, aqueous liquid cosmetic composition formation is the most prevalent. Similarly, in abstract topic ID 1 and 2 (#9 and 15-Annexure II) related to the preparation of pharmaceutical salts for drugs have a greater number of patents under these. It is well known that though pyLDAvis and Gensim are the same, they do not use the same ID numbers for topics. The right panel with the horizontal bar graph indicates the most representative words of each topic. While the blue bars represent the overall frequency of each word in the corpus, red bars represent the estimated number of times a given term appeared in a given topic. Relationships among the topics in Figures 4a and 4b are also depicted through bubbles, which are away from each other, indicating they are unique and do not have any overlap with other topics. However, the distance between two bubbles indicates the semantic distance between the topics (intertopic distance), lesser the distance is, the more they are interrelated (similar) to

**Table 1: Occurrences of top twenty areas of Indian healthcare related patents as per IPC classes.**

| IPC Class | IPC Sub-Classes | N | Description of IPC Subjects |
|---|---|---|---|
| A61K- PREPARATIONS FOR MEDICAL, DENTAL OR TOILETRY PURPOSES | A61K 9/20 | 727 | Medicinal preparations characterised by special physical form - Pills, lozenges or tablets. |
| | A61K 31/00 | 373 | Medicinal preparations containing organic active ingredients. |
| | A61K 9/16 | 274 | Medicinal preparations characterised by special physical form - Agglomerates; Granulates; Microbeadlets. |
| | A61K 36/00 | 251 | Medicinal preparations of undetermined constitution containing material from algae, lichens, fungi or plants, or derivatives thereof, e.g. traditional herbal medicines. |
| | A61K 9/28 | 234 | Medicinal preparations characterised by special physical form - Dragees; Coated pills or tablets. |
| | A61K 9/14 | 210 | Medicinal preparations characterised by special physical form - Particulate form, e.g. powders. |
| | A61K 9/48 | 200 | Medicinal preparations characterised by special physical form - Preparations in capsules, e.g. of gelatin, of chocolate. |
| | A61K 45/06 | 189 | Medicinal preparations characterised by Mixtures of active ingredients without chemical characterisation, e.g. antiphlogistics and cardiac. |
| A61P - SPECIFIC THERAPEUTIC ACTIVITY OF CHEMICAL COMPOUNDS OR MEDICINAL PREPARATIONS | A61P 35/00 | 529 | Antineoplastic agents. |
| | A61P 29/00 | 222 | Non-central analgesic, antipyretic or anti-inflammatory agents, e.g. antirheumatic agents; Non-steroidal anti-inflammatory drugs. |
| | A61P 31/04 | 203 | Antibacterial agents. |
| | A61P 3/10 | 201 | Drugs for disorders of the metabolism for hyperglycaemia, e.g. antidiabetics. |
| | A61P 25/00 | 162 | Drugs for disorders of the nervous system. |
| A61Q- SPECIFIC USE OF COSMETICS OR SIMILAR TOILETRY PREPARATIONS | A61Q 19/00 | 201 | Preparations for care of the skin. |
| | A61Q 5/12 | 147 | Preparations containing hair conditioners. |
| | A61Q 5/02 | 141 | Preparations for cleaning the hair |
| | A61Q 19/10 | 140 | Washing or bathing preparations. |
| C07D - ORGANIC CHEMISTRY - HETEROCYCLIC COMPOUNDS | C07D 487/04 | 180 | Heterocyclic compounds containing nitrogen atoms as the only ring hetero atoms in the condensed system, at least one ring being a six-membered ring with one nitrogen atom- Ortho-condensed system. |
| | C07D 471/04 | 171 | Heterocyclic Compounds containing nitrogen atoms as the only ring hetero atoms in the condensed system -Ortho-condensed system. |
| | C07D 401/12 | 152 | Heterocyclic compounds containing two or more hetero rings, having nitrogen atoms as the only ring hetero atoms, and linked by a chain containing hetero atoms as chain links. |

N= number of patents

**Table 2: Participation of Triple Helix institutions in the medical innovation of India.**

| University | No. | Industry | No. | Government | No. |
|---|---|---|---|---|---|
| Indian Institute of Science | 46 | Hindustan Lever Ltd. | 746 | Council of Scientific and Industrial Research (CSIR) | 272 |
| Indian Institute of Technology Madras | 28 | Ranbaxy Laboratories Ltd. | 494 | Department of Biotechnology | 27 |
| Indian Institute of Technology Bombay | 28 | Cipla Ltd. | 213 | National Institute of Immunology | 26 |
| Indian Institute of Technology Delhi | 20 | Cadila Healthcare Ltd. | 207 | Jawaharlal Nehru Centre for Advanced Scientific Research | 29 |
| Indian Institute of Technology Kanpur | 9 | Sun Pharmaceutical Industries Ltd. | 176 | Indian Council of Medical Research | 17 |
| Amrita Vishwa Vidyapeetham University | 7 | Wockhardt Ltd. | 156 | International Centre for Genetic Engineering and Biotechnology | 17 |
| University of Calcutta | 7 | Dr. Reddy's Laboratories, Inc. | 144 | All India Institute of Medical Sciences | 11 |
| University of Delhi | 6 | Cellixbio Private Ltd. | 93 | Healthcare Technology Innovation Centre | 10 |
| University of Mysore | 5 | Hetero Research Foundation | 81 | Institute of Life Sciences | 7 |
| University of Hyderabad/ Panjab University | 4 | Piramal Enterprises Ltd. | 76 | Rajiv Gandhi Centre for Biotechnology | 6 |

No.=Number of Patents

each other. Therefore, it may be fair to say that abstract topics of our dataset show greater semantic relations among them than title topics, as the distance between the topics in the abstract is nearer. Topic ID 1 (Annexure II # 13-related to cancer drug) in the abstract has a lesser distance with topic ID 4 (Annexure II # 9-related to Heterocyclic drugs) and topic 4 is nearer to topic ID 3 (Annexure II # 4-herbal extracted anti-viral agents). Similarly, title topic ID 1 (Annexure II # 20) has a lesser distance with topic ID 5 (topic no. 15) and topic ID 3 (topic no. 11).

It was of our interest to know whether the LDA topics for titles and abstract serve as a complement to each other or serve as a supplement to one another. To comply, we adopted an alternative approach for implementing topic information based on sub-words into the most accepted language models, namely Sentence-BERT (SBERT). This model helps us in the analysis of both Semantic Textual Similarity and Semantic Similarity Detection between the topics.[29] We found evidence that to determine the most similar sentence in a collection and find cosine likeness within, SBERT can be utilized.[30] This model can also be used to understand how semantically related the latent topics are Dang *et al*., 2022.[31] Therefore, we used the SBERT cosine similarity score, ranging from 0 to 1, scoring 0 for pairs that are dissimilar and 1 for pairs that are similar. As shown in Table 3, the title and abstract topics have an overall semantic similarity score of 0.611001. Almost 15% of the topics have an inter-topic similarity score of almost 50% or above. Remaining 75% of topics have 29 to 45% of semantic similarity.

From Table 3, it is evident that only 15% title topics have almost 50% or more, similarly with abstract topics. Now to explain which topics of title have more semantic relation with the abstract, attempts were made to explore similarly with the human-intervened umbrella topics. The findings of the analysis in Table 4, ranked according to decreasing value of documents per umbrella topic, shows that patents related to pharmaceutical preparation of drug salt, anti-cancer drugs development, preparation of drugs for anti-microbial infection, herbal drug formulation, preparation of compound for cosmetic use, preparation of natural polymers, pharmaceutical use of nanoparticles are common topics in both the two fields. However, patents related to the purification of water, preparation of amorphous compounds, plant based anti-diabetic drug development are common in title umbrella topics but are uncommon in the abstract. Similarly, patents related to wound healing, medical equipment like beds, surgery items, and cardiac tissue engineering are visible in the abstract umbrella topic but unavailable in the title umbrella topics. Comparing the IPC class of titles and abstracts, it was noticed that IPC class for titles is mostly up to subclass or groups. However, the IPC classes of the abstracts are mostly up to the sub-groups, which means abstract topics are more precise than the title topics. It is also worth noting that all the predominant IPC classes that are identified in Table 1 are also identified with the corresponding IPC topics of titles and abstracts.

Finally, while adjusting a record number of titles and abstracts by topic through LDA model with the original record number derived through IPC class, we observed similar results in terms

of university-industry-government participation. A significant number of industry than university or government participation for inventing patents. This means that industrial organizations are more focused on the expansion of technologies and developing related technologies for their customers. In the medical sectors, commercial multinational companies, including Dr. Reddy's Lab, Cipla, Sun Pharma, and Lupin have taken considerable initiatives to develop new drugs for human needs or commercial gain, while government-funded sectors like CSIR and DBT are putting their interest towards expansion and enhancing existing technologies related to cancer research or target drug delivery methods.

Table 5 presents a comparison between the manual IPC classifications and the machine-generated LDA, title, and abstract topics. Because IPC classes are assigned based on the technical characteristics and functional attributes of each invention, relying solely on patent titles often makes accurate classification challenging. The machine-generated topics nevertheless show clear and meaningful alignment with key IPC categories. The most frequent IPC class, A61K (Preparations for Medical, Dental, or Toilet Purposes), corresponds closely with LDA title topics such as "crystalline pharmaceutical compositions," "creams," and "herbal formulations," as well as abstract topics that reflect greater chemical specificity, including "pharmaceutical preparation of solifenacin or a salt" and "heterocyclic compounds." The next prominent IPC class, A61P (Specific Therapeutic Activity of Chemical Compounds or Medicinal Preparations), also aligns with machine-derived topics related to indole-based compounds, pesticides and fungicides, and pyrimidine-derived pharmaceutical salts. Similarly, IPC class A61Q (Specific Use of Cosmetics or Similar Toilet Preparations) matches machine-generated themes focused on herbal skin-whitening formulations and cosmetic-grade surfactants. Therefore, abstract-based topics tend to correspond more closely with IPC classifications because they capture structural and chemical details such as heterocyclic or pyrimidine derivatives. Title-based topics highlight higher-level applications or product names, including drug capsules or protein kinase inhibitors. These patterns indicate that the machine-generated topic models effectively capture the major IPC categories represented in the patent dataset.

## DISCUSSION

Since the intention of this study is to perceive knowledge on which machine algorithms can supplement the thinking capacity of the human brain, we have tested the same set of data under two situations. To gain knowledge on the major topics for the published patents of a country like India and for a domain like healthcare, we studied the human-assigned IPC-based class numbers of published patents and grouped them under major areas/topics of innovation based on their occurrence. The title and abstract fields of these patents were used in LDA model to extract machine-generated 20 topics and then compared to understand how far topics extracted from both processes are semantically similar or not. As patents are the invention of innovators, our intention was also to identify whether the major

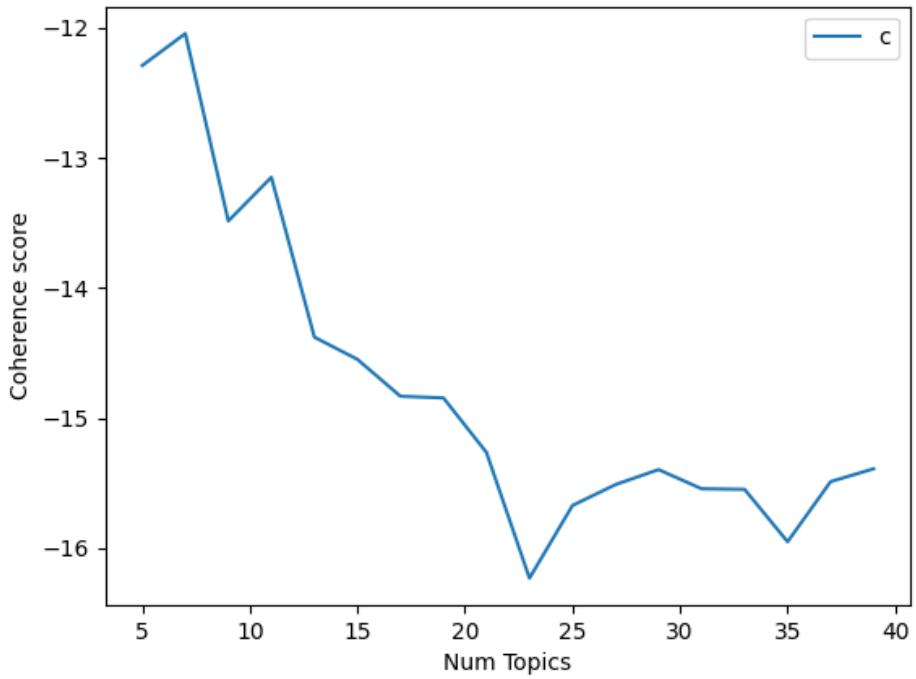


**Figure 2a:** Coherence matrix for Titles.

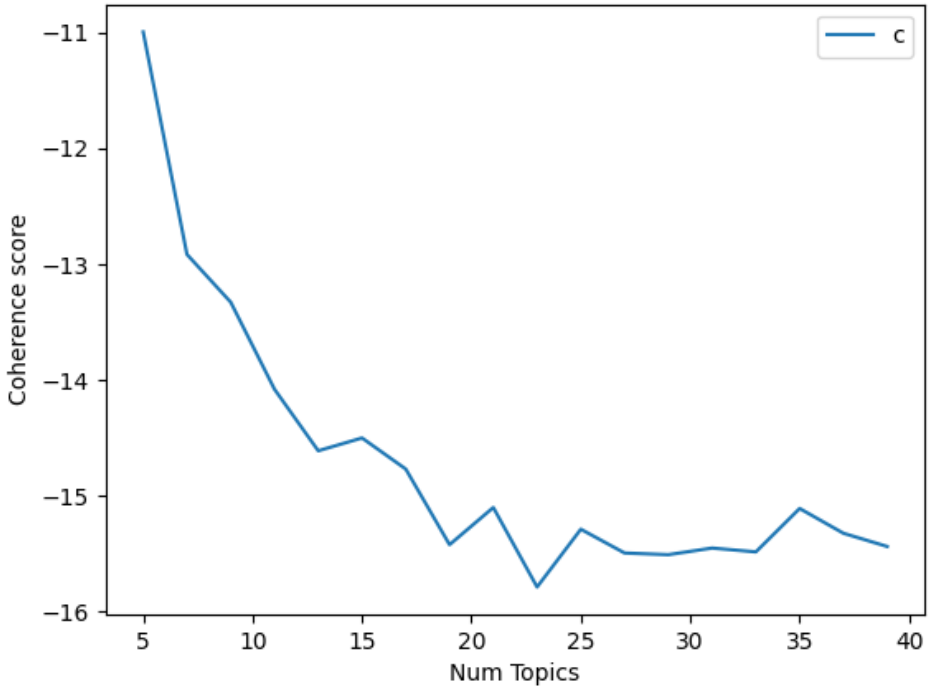


**Figure 2b:** Coherence matrix for abstracts.

**Table 3: Similarity score of title topics and abstract topic through SBERT.**

| Title Topic | Top three Abstract topic with similarity score | | | | | |
|---|---|---|---|---|---|---|
| | Topic No | Similarity Score | Topic No | Similarity Score | Topic No | Similarity Score |
| 0 | 8 | 0.6405 | 9 | 0.5852 | 4 | 0.5480 |
| 1 | 9 | 0.5972 | 19 | 0.4871 | 14 | 0.4674 |
| 2 | 11 | 0.5315 | 16 | 0.5173 | 15 | 0.5114 |
| 3 | 18 | 0.7301 | 19 | 0.6167 | 1 | 0.5599 |
| 4 | 13 | 0.6382 | 9 | 0.6057 | 11 | 0.5842 |
| 5 | 16 | 0.5860 | 15 | 0.5651 | 9 | 0.5513 |
| 6 | 4 | 0.6739 | 18 | 0.5452 | 0 | 0.5297 |
| 7 | 15 | 0.5883 | 11 | 0.5503 | 9 | 0.5494 |
| 8 | 4 | 0.6531 | 18 | 0.5522 | 19 | 0.5504 |
| 9 | 9 | 0.6067 | 15 | 0.5354 | 11 | 0.4931 |
| 10 | 14 | 0.6070 | 9 | 0.5019 | 11 | 0.4957 |
| 11 | 19 | 0.7582 | 15 | 0.7147 | 4 | 0.6158 |
| 12 | 11 | 0.6928 | 9 | 0.6221 | 15 | 0.6034 |
| 13 | 15 | 0.5823 | 18 | 0.5717 | 17 | 0.5158 |
| 14 | 16 | 0.5203 | 19 | 0.5020 | 1 | 0.4967 |
| 15 | 0 | 0.7503 | 15 | 0.6542 | 11 | 0.5739 |
| 16 | 1 | 0.6362 | 9 | 0.5529 | 15 | 0.5447 |
| 17 | 12 | 0.5570 | 9 | 0.5473 | 13 | 0.5443 |
| 18 | 5 | 0.6157 | 10 | 0.5727 | 4 | 0.5401 |
| 19 | 8 | 0.6092 | 17 | 0.5720 | 3 | 0.5342 |

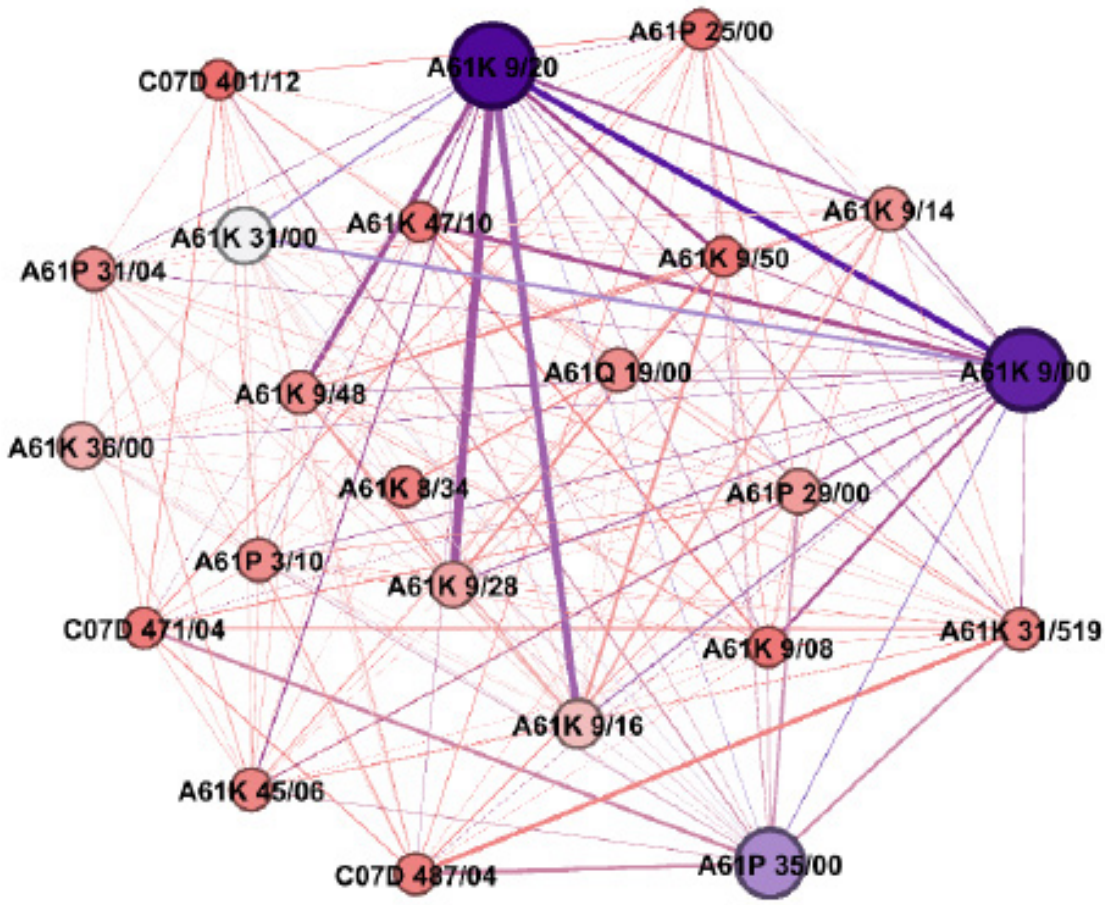


**Figure 3:** Extent of Inter-relation between significant sub-class of Patents related to medical innovation.

players of innovation for both the processes i.e., human-assigned and machine-derived, are the same or different.

It is well understood that the syntax of writing the title of a patent is different from that of a scientific research article. Some of the words in the title of the patent have no semantic value with the meaning of the thought contents in the title. Sometimes, two or more patents have been assigned the same title, and some patents have only a single word in the title, resulting in the generated document feature matrix being too sparse. This led to the difficulty for the LDA model to cluster some documents with different subject contents but the same word usage and appropriate word frequency into a class. Therefore, for identifying latent topics, along with titles, the abstract field should also be considered.

In our study, we have chosen twenty topics from both title and abstract based on almost the same u_mass coherence score. However, we observed that topics that are identified through the LDA Gensim model for titles are different than abstracts for the same corpus data. Though overall semantic similarity was observed in almost 60% between tile and abstract topics, almost 40% topics are not similar to each other. Only 15% topics of both titles and abstracts have more than 50% cosine similarity score, which indicates that titles and abstracts are different entities of a corpus of text. Therefore, they are not complementary to each other. For a machine learning approach, it is therefore better to choose both the title and abstract to measure the latent topic or trends of a subject comprehensively.

During analysis through human-assigned class numbers, it was observed that drug development is the major area of healthcare innovation, and in an industrial setup, MNCs are the major players of innovation in the healthcare sector. This is because pharmaceutical companies want to recoup investment by charging a monopoly price for their products European Commission, 2009.[32] As India became a part of GATT and signed TRIP's agreement, drug patents as product patents for 20 years have become more efficient. Following this, cosmetic production for human necessities is the second top area of innovation, for which industries are also the major players.

**Table 4: Latent topics from titles and abstracts of patents.**

| TN | Umbrella Title Topic | NPT | CIPC | T3O | TN | Umbrella Abstract Topic | NPT | CIPC | T3O |
|---|---|---|---|---|---|---|---|---|---|
| 15 | Crystalline pharmaceutical composition of drugs, creams etc. | 1779 | A61K 31/00, A01N 43/00 | RANBAXY, CSIR, CADILA | 9 | Pharmaceutical preparation of solifenacin or a salt - Heterocyclic compounds | 2323 | C07D453/02 | WOCKHARDT, DRREDDY, GLENMARK |
| 6 | Process of making herbal compositions | 1075 | A61K 9/00; A61K 36/185, C07K 14/705 | CSIR, DRREDDY, HUL | 15 = 1 | Pharmaceutical salts of pyrimidine derivatives and method of treating disorders | 1841 | C07D 403/04 C07D 403/00 | GLENMARK, CIPLA, CELLIXBIO |
| 0 | Natural polymer to prepare capsule for drugs | 1001 | A61K 9/00, A61K 9/48 | SUN, DRREDDY, NII | 4 | Herbal extracted anti-viral agents/ Rotavirus vaccine composition | 1290 | A61K 39/15, A61K 39/12, A61K 9/00 | B.BIO, HUL, UPL |
| 19 | Protein kinase inhibitor | 830 | A01N 63/02, C07D 401/00 | LUPIN, AURIGENE, CADILA | 1 | Aqueous liquid cosmetic composition-grade surfactants | 918 | A61Q 1/04, A61K 8/891 A61K 8/39 | SUN, HUL, LUPIN |
| 8 | Herbal preparation of skin whitening compounds | 805 | A61K 9/00; A61F2/06 | CSIR, PEL, HUL | 11 | Process for preparation of pharmaceutically accepted salts | 846 | C07D 487/00 | DRREDDY, RANBAXY, MSD |
| 11 | Method of treating/ preventing inhibiting respiratory and respiratory virus | 722 | A61K 31/045, A61K 31/07 | GLENMARK, CIPLA, WOCKHARDT | 13 | Anti-Inflammatory Agents for Cancer Therapy | 818 | A61J 3/00, A61K 31/00 | DRREDDY, CIPLA, CSIR, |
| 13 | Plant extract methods and use for treatment of diabetics and other diseases | 650 | A61P 3/10, A61M 5/32, A61K 36/54 | AVESTHAGEN, CIPLA, LUPIN | 0 | Cationic surfactants in hair conditioning formulation | 438 | A61Q5/12, A61K 7/06 | HUL, CSIR, DRREDDR |
| 12 | Formulation of oral dosages of amorphous compounds | 649 | A61K 9/00, A61K31/422, C07D403/14 | CADILA, DRREDDY, RANBAXY | 3 | Sensors for sensing a plurality of parameters | 415 | G01D 5/2241, G01D 5/12 | TCS,DBT, HTIC |
| 7 | Method of preparing cosmetic compositions, therapeutic management of obesity | 617 | A61K 8/00, A61Q 5/12 A61K 31/12 | HUL, RANBAXY, CIPLA | 7 | Hospital medical equipment/Devices | 392 | G16H | DBT, MERIL, LUPIN |
| 9 | Pharmaceutically acceptable indole derivatives preparation and use as pesticides, fungicides, insecticides | 611 | C07D 401/14, A01N 63/00, C07D209/08 | UPL, INSECTICID, LUPIN | 8 | Protein-based Therapeutics delivery agents | 229 | A61K | CSIR, SERUM, NII |
| 5 | Preparation of nano-coating antimicrobial compound | 337 | A01N 25/00, A61P 31/04, A61K 31/44 | RANBAXY, DRREDDY, CADILA | 17 | Preparation of therapeutic peptide for clinical purpose | 194 | C07K 1/00, AK61K 38/00 | SUN, MYLAN, BIOCON |

| TN | Umbrella Title Topic | NPT | CIPC | T3O | TN | Umbrella Abstract Topic | NPT | CIPC | T3O |
|---|---|---|---|---|---|---|---|---|---|
| 17 | Targeted drug delivery methods and devices | 274 | A61K 9/00, A61K 49/00, A61M 5/14 | CSIR, POLYMED, MERIL | 19 | Medical preparation of sodium alginate for wound healing | 191 | A61L 26/00, A61K 47/00 | CADILA,DRREDDY, HUL |
| 16 | Composition and use of surface modified meta nanoparticples | 270 | A61K 33/24, A01N 59/00, C07D 309/28 | SUN, DRREDDY, MSD | 18 | Silicon-based devices for delivery of therapeutic agents, Herbal preparation for cosmetics. | 164 | A61K 9/00, A 61K 36/00, A61P | HUL, LUPIN, DRREDDY |
| 1 | Preparation of anti-inflammatory pharmaceutical compositions/ salts | 225 | A61K 31/00; A61K 36/00 A61P 29/00 | CSIR, NCCS, RANBAXY | 2 | Cycloalkyl Amine Compounds (Heterocyclic) for treatment of hyperproliferative diseases | 154 | C07D 295/037, C07D 295/02 | CSIR, SUN, GLENMARK |
| 18 | Molecular complex preparation and use | 217 | A01N 37/00, A61K 47/30; A61K 51/00 | RANBAXY, CIPLA, CSIR | 10 | Storing device for biological materials/ cavity resonator | 143 | A61B G01N 29/022 | POLYMED, FORUS, LUPIN |
| 3 | Water-in-oil multiple emulsions and methods for use in adjuvants, pharmaceuticals, cosmetics, foods | 214 | A61K 31/00, A61Q 19/00 | HUL, RANBAXY, IITB | 5 | Cardiac tissue engineering/ Developing functional tissues for cardiac regeneration | 120 | A61L 27/3839 A61K 9/0024 | MERIL, IITM, SERUM |
| 10 | Polymorph drugs manufacture for treatment of infection | 213 | A61k 31/00, C07C 231/24, C07H 17/08 | RANBAXY, GENERICS, CADILA | 14 | Components for targeted drug delivery system | 114 | A61K 47/00 | DRREDDY, CSIR, SHILPA |
| 4 | Anticancer peptide drug development | 182 | A61K 38/00, A61K 45/06, A61K 39/198, C07K 14/00 | CSIR, SUN, CELAGENEX | 16 | Nanocomposite inhibitors for antibacterial', 'antifungal', activity | 106 | A01N 59/16, A61P 31/04 | CSIR, RANBAXY, DRREDDY |
| 2 | Methods for purification of amorphous water | 121 | C02F 1/583, C01B 33/46, A01N 65/00, B65D 81/00 | DRREDDY, RANBAXY, HUL | 12 | Device based on Lock and release mechanisms for trans-catheter implantable devices/ Surgical instruments | 105 | A61B 17/346, A61B 17/3468 | POLYMED, DBT,MERIL |
| 14 | Preparation of plant based anti-diabetic, antiviral nono-formulations | 99 | A01N 63/04, A61K 36/00 A61K 39/00 | CSIR, BIOLOGICALE, DRREDDY | 6 | Hydrogen-bonding enhanced Nano-carrier drug delivery in vascular system | 67 | A61K 49/0002, A61K 9/1271, A61K 8/73 | CSIR, MERIL, MSD |

NPT= Number of Patent Tile, CIPC= Corresponding IPC Class, T3O= Top 3 U-I-G Organizations, TN=Topic numberAURIGENE- Aurigene Discovery Technologies Ltd.; AVESTHAGEN - AvesthaGengraine Technologies Pvt. Ltd.; B. BIO-Bharat Biotech International Ltd.; BIOCON-Biocon Ltd.; BIOLOGICALE-Biological E Ltd.; CADILA-Cadila Healthcare Ltd.; CELAGENEX-Celagenex Research (India) Pvt. Ltd.; CELLIXBIO-Cellixbio Private Ltd.; CIPLA-Cipla Ltd.; CSIR-Council Of Scientific and Industrial Research; DBT-Secretary, Department Of Biotechnology; DRREDDR-Dr. Reddy's Laboratories Ltd.; FORUS-Forus Health Pvt. Ltd.; GENERICS-Generics [Uk] Ltd.; GLENMARK-Glenmark Pharmaceuticals Ltd; HTIC-Healthcare Technology Innovation Centre; HUL-Hindustan Lever Ltd.; IITB-Indian Institute Of Technology Bombay; IITM-Indian Institute Of Technology Madras; LUPIN-Lupin Ltd.; MERIL-Meril Life Sciences Pvt Ltd; MSD-MsdWellcome Trust Hilleman Laboratories Pvt. Ltd.; MYLAN-Mylan Laboratories Ltd.; NCCS-National Centre For Cell Science; NII-National Institute Of Immunology; PEL-Piramal Enterprises Ltd.; POLYMED-Poly MedicureLtd.; RANBAXY-Ranbaxy Laboratories Ltd.; SERUM-Serum Institute Of India Ltd.; SHILPA-Shilpa Medicare Ltd.; SUN-Sun Pharmaceutical Industries Ltd.; TCS-Tata Consultancy Services Ltd.; UPL-UplLtd.; WOCKHARDT-Wockhardt Ltd.

**Table 5: Comparative analysis of man-made (IPC Subjects) topic and machine generated (Title topic and Abstract Topic) patent topic.**

| Title Topic | IPC Subjects | Abstract Topic |
|---|---|---|
| Crystalline pharmaceutical composition of drugs, creams etc. | Medicinal preparations characterized by special physical form - Pills, lozenges or tablets. | Pharmaceutical preparation of solifenacin or a salt - Heterocyclic compounds. |
| Process of making herbal compositions | Medicinal preparations containing organic active ingredients. | Pharmaceutical salts of pyrimidine derivatives and method of treating disorders. |
| Natural polymer to prepare capsule for drugs | Medicinal preparations characterised by special physical form - Agglomerates; Granulates; Microbeadlets. | Herbal extracted anti-viral agents/ Rotavirus vaccine composition. |
| Protein kinase inhibitor | Medicinal preparations of undetermined constitution containing material from algae, lichens, fungi or plants, or derivatives thereof, e.g. traditional herbal medicines. | Aqueous liquid cosmetic composition-grade surfactants. |
| Herbal preparation of skin whitening compounds | Medicinal preparations characterised by special physical form - Dragees; Coated pills or tablets. | Process for preparation of pharmaceutically accepted salts. |
| Method of treating/preventing inhibiting respiratory and respiratory virus | Medicinal preparations characterised by special physical form - Particulate form, e.g. powders. | Anti-Inflammatory Agents for Cancer Therapy. |
| Plant extract methods and use for treatment of diabetics and other diseases | Medicinal preparations characterised by special physical form - Preparations in capsules, e.g. of gelatin, of chocolate. | Cationic surfactants in hair conditioning formulation. |
| Formulation of oral dosages of amorphous compounds | Medicinal preparations characterised by Mixtures of active ingredients without chemical characterisation, e.g. antiphlogistics and cardiac. | Sensors for sensing a plurality of parameters. |
| Method of preparing cosmetic compositions, therapeutic management of obesity | Antineoplastic agents | Hospital medical equipment/Devices. |
| Pharmaceutically acceptable indole derivatives preparation and use as pesticides, fungicides, insecticides | Non-central analgesic, antipyretic or antiinflammatory agents, e.g. antirheumatic agents; Non-steroidal antiinflammatory drugs. | Protein based Therapeutics delivery agents. |
| Preparation of nano-coating antimicrobial compound | Antibacterial agents. | Preparation of therapeutic peptide for clinical purpose. |
| Targeted drug delivery methods and devices | Drugs for disorders of the metabolism for hyperglycaemia, e.g. antidiabetics. | Medical preparation of sodium alginate for wound healing. |
| Composition and use of surface modified meta nanoparticples | Drugs for disorders of the nervous system. | Silicon-based devices for delivery of therapeutic agents, Herbal preparation for cosmetics. |
| Preparation of anti-inflammatory pharmaceutical compositions/ salts | Preparations for care of the skin. | Cycloalkyl Amine Compounds (Heterocyclic) for treatment of hyperproliferative diseases. |
| Molecular complex preparation and use | Preparations containing hair conditioners. | Storing device for biological materials/ cavity resonator. |
| Water-in-oil multiple emulsions and methods for use in adjuvants, pharmaceuticals, cosmetics, foods | Preparations for cleaning the hair. | Cardiac tissue engineering/ Developing functional tissues for cardiac regeneration. |

| Title Topic | IPC Subjects | Abstract Topic |
|---|---|---|
| Polymorph drugs manufacture for treatment of infection | Washing or bathing preparations. | Components for targeted drug delivery system. |
| Anticancer peptide drug development | Heterocyclic compounds containing nitrogen atoms as the only ring hetero atoms in the condensed system, at least one ring being a six-membered ring with one nitrogen atom- Ortho-condensed system. | Nanocomposite inhibitors for antibacterial', 'antifungal', activity. |
| Methods for purification of amorphous water | Heterocyclic Compounds containing nitrogen atoms as the only ring hetero atoms in the condensed system -Ortho-condensed system. | Device based on Lock and release mechanisms for trans-catheter implantable devices/ Surgical instruments. |
| Preparation of plant based anti-diabetic, antiviral nono-formulations | Heterocyclic compounds containing two or more hetero rings, having nitrogen atoms as the only ring hetero atoms, and linked by a chain containing hetero atoms as chain links. | Hydrogen-bonding enhanced Nano-carrier drug delivery in vascular system. |

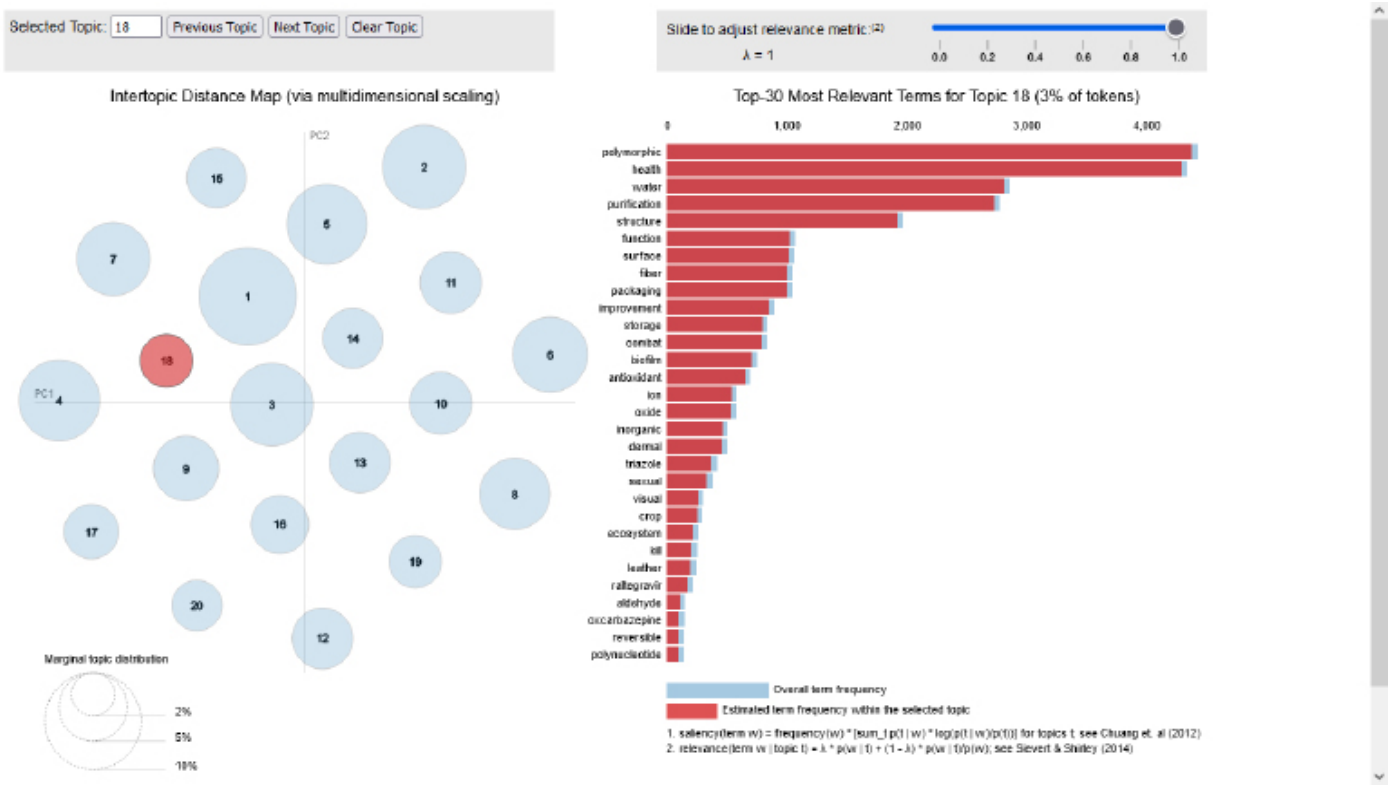


**Figure 4a:** pyLDAvis Graph for Title terms.

While comparing results of SBERT-based machine extracted results with human-intervened machine results, we found that though the machine can project the best possible similarity, for choosing umbrella topics, still human intervention is necessary. Umbrella title topics like 'Crystalline pharmaceutical composition of drugs, creams, etc.' were found in 17% documents, followed by 'Process of making herbal compositions' in 10% documents. Similarly, in the abstract umbrella topic, Heterocyclic compounds based pharmaceutical preparation of solifenacin or a salt was found in almost 23% documents, followed by Pharmaceutical salts of pyrimidine derivatives and method of treating disorders in 18% documents. It was seen that the topics extracted from man-machine interaction are more precise, deep-rooted (sub-classes) than simply human-assigned topics through IPC class. Therefore, based on our analysis it may not be unfair to say that the man-machine interface can efficiently assign topics similar to those assigned manually.

## CONCLUSION

With the abundance of unstructured data and language fragments, unsupervised approaches like topic modelling are employed to successfully extract topic keywords from a document through a semantic generalization of the document's content. By analyzing

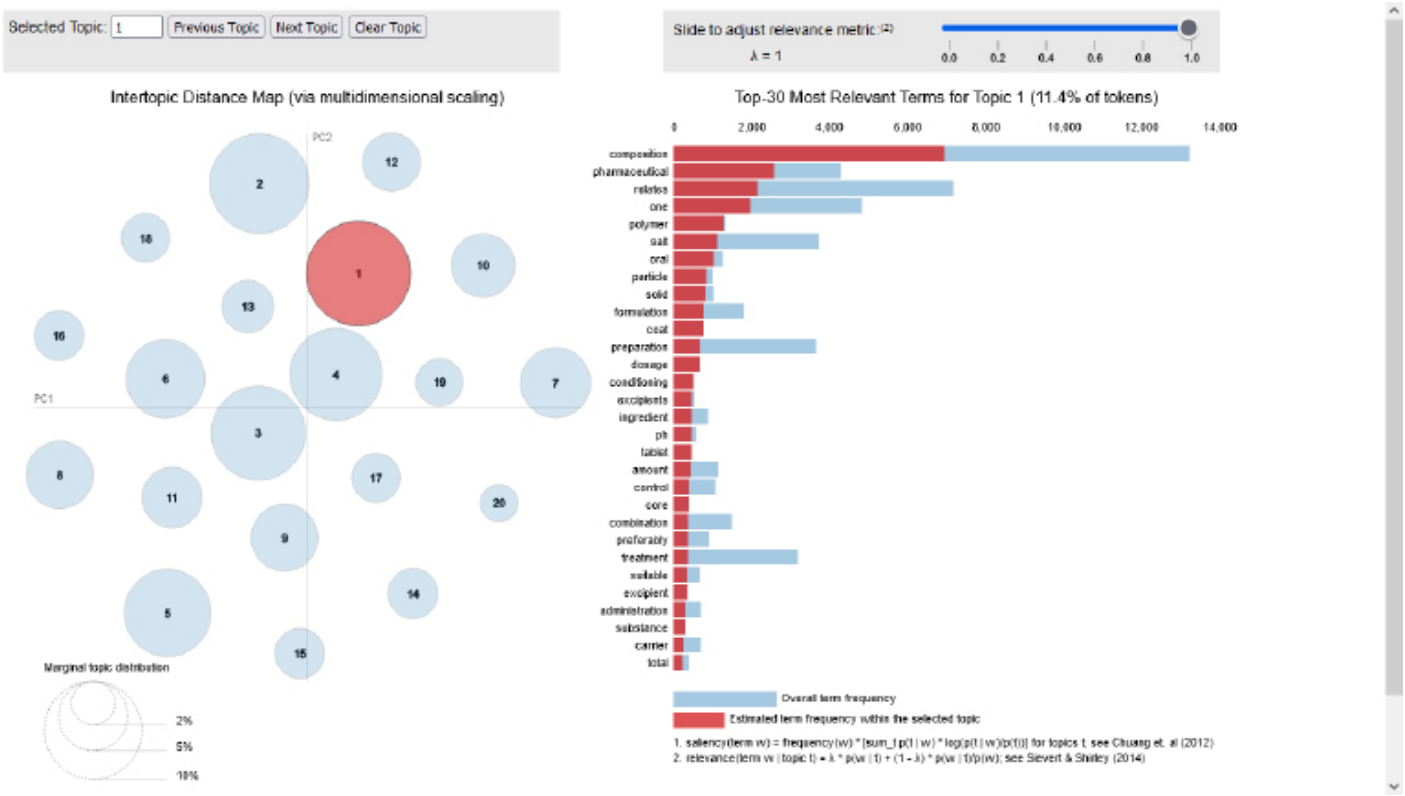


**Figure 4b:** pyLDAvis Graph for Title terms.

the textual content of the documents, topic modelling algorithms assign probabilities to words' association with specific topics, allowing them to generate a comprehensive representation of the underlying themes present in the collection. Although LDA can generate topic modeling from a variety of large collections, it is not very suitable for short-length documents. Another limitation includes that LDA topic modeling needs a predefined number of topics, which may affect the actual result. In this study, we found that LDA model-based topic modelling can be used as an effective means for identifying latent topics of a text corpus. However, we believe that, along with the title, the abstract field is also essential for identifying the latent topics of patents. To better verify the feasibility of the findings proposed in this study, a further analysis is needed with patent data from other fields. To our belief, this study will contribute to the healthcare sector in identifying the existing domains of healthcare innovation in India so that stakeholders can channelize their resources in the right direction. Furthermore, the empirical evidence gained through LDA analysis will enable us to understand which title and abstract fields can discover important latent topics to predict the trends of innovation of a country under a specific domain.

## ABBREVIATIONS

**LDA:** Latent Dirichlet Allocation; **IPC:** International Patent Classification; **WIPO:** World Intellectual Property Organization; **PCT:** Patent Cooperation Treaty; **USPTO:** United States Patent and Trademark Office; **EPO:** European Patent Office; **DPMA:** German Patent and Trademark Office; **CIPO:** Canadian Intellectual Property Office; **IT:** Information Technology; **R&D:** Research and Development; **GDP:** Gross Domestic Product; **GATT:** General Agreement on Tariffs and Trade; **TRIPS:** Trade-Related Aspects of Intellectual Property Rights; **NLTK:** Natural Language Toolkit; **SBERT:** Sentence Bidirectional Encoder Representations from Transformers; **AI:** Artificial Intelligence; **ML:** Machine Learning; **NLP:** Natural Language Processing; **CSIR:** Council of Scientific and Industrial Research; **DBT:** Department of Biotechnology; **ICMR:** Indian Council of Medical Research; **IISc:** Indian Institute of Science; **IIT:** Indian Institute of Technology; **HUL:** Hindustan Unilever Limited; **RAM:** Random Access Memory; **CPU:** Central Processing Unit; **mRNA:** Messenger Ribonucleic Acid.

## CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

## AUTHOR CONTRIBUTION

Conceptualization: Debasis Majhi, Methodology: Bhaskar Mukherjee, Validation: Debasis Majhi, Formal Analysis: Bhaskar Mukherjee, Writing - Draft preparation: Priya Tiwari.

## REFERENCES

1. Cockburn I, Long G. The importance of patents to innovation: updated cross-industry comparisons with biopharmaceuticals. Expert Opin Ther Pat. 2015;25(7):739-42. doi: 10.1517/13543776.2015.1040762, PMID 25927945.

2. WIPO guide to using patent information; 2021. Available from: https://www.wipo.int/edocs/pubdocs/en/wipo-pub-rn2021-1e-en-wipo-guide-to-using-patent-information.pdf.
3. Global innovative index; 2019. Creating Healthy Lives – the Future of Medical Innovation. Available from: https://www.wipo.int/wipo_magazine/en/2019/04/article_0001.html.
4. Flessa S, Huebner C. Innovations in health care-A conceptual framework. Int J Environ Res Public Health. 2021;18(19):10026. doi: 10.3390/ijerph181910026, PMID 34639328.
5. Mayfield DL. Medical patents and how new instruments or medications might be patented. Mo Med. 2016;113(6):456-62. PMID 30228529.
6. Giglio C, Vocaturo GS, Palmieri R. Patent acquisitions in the healthcare industry: an analysis of learning mechanisms. Int J Environ Res Public Health. 2023;20(5):4100. doi: 10.3390/ijerph20054100, PMID 36901110.
7. Guide to the international patent classification; 2023. Available from: https://www.wipo.int/publications/en/details.jsp?id=4656andplang=EN.
8. Mazumdar-Shaw K. Leveraging affordable innovation to tackle India's healthcare challenge. IIMB Manag Rev. 2018;30(1):Article 1. doi: 10.1016/j.iimb.2017.11.003.
9. Chae S, Gim J. A study on trend analysis of applicants based on patent classification systems. Information. 2019;10(12):Article 12. doi: 10.3390/info10120364.
10. Tsapatsoulis N, Partaourides H, Christodoulou C, Djouvas C. Quo vadis computer science? The topics of the influential papers during the period 2014-2021 The topics of the influential papers during the period IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech); 2022. p. 1-8. doi: 10.1109/DASC/PiCom/CBDCom/Cy55231.2022.9927789.
11. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3(January):993-1022.
12. San Torcuato M, Bautista-Puig N, Arrizabalaga O, Méndez E. Tracking openness and topic evolution of COVID-19 publications January 2020-March 2021: comprehensive bibliometric and topic modeling analysis. J Med Internet Res. 2022;24(10):e40011. doi: 10.2196/40011, PMID 36190742.
13. Basilio MP, Pereira V, Oliveira MW, de. Knowledge discovery in research on policing strategies: an overview of the past fifty years. J Modell Manag. 2021;17(4):1372-409. doi: 10.1108/JM2-10-2020-0268.
14. Chen Y, Ghosh J, Bejan CA, Gunter CA, Gupta S, Kho A, *et al*. Building bridges across electronic health record systems through inferred phenotypic topics. J Biomed Inform. 2015;55:82-93. doi: 10.1016/j.jbi.2015.03.011, PMID 25841328.
15. Erosheva E, Fienberg S, Lafferty J. Mixed-membership models of scientific publications. Proc Natl Acad Sci U S A. 2004;101 Suppl 1:Article suppl_1. doi: 10.1073/pnas.0307760101, PMID 15020766.
16. Ali I, Kannan D. Mapping research on healthcare operations and supply chain management: a topic modelling-based literature review. Ann Oper Res. 2022;315(1):29-55. doi: 10.1007/s10479-022-04596-5, PMID 35382453.
17. Lu HM, Wei CP, Hsiao FY. Modeling healthcare data using multiple-channel latent Dirichlet allocation. J Biomed Inform. 2016;60:210-23. doi: 10.1016/j.jbi.2016.02.003, PMID 26898516.
18. Fairie P, Zhang Z, D'Souza AG, Walsh T, Quan H, Santana MJ. Categorising patient concerns using natural language processing techniques. BMJ Health Care Inform. 2021;28(1):e100274. doi: 10.1136/bmjhci-2020-100274, PMID 34193519.
19. Yang S, Bian J, Sun Z, Wang L, Zhu H, Xiong H, *et al*. Early detection of disease using electronic health records and fisher's Wishart discriminant analysis. Procedia Comput Sci. 2018;140:393-402. doi: 10.1016/j.procs.2018.10.299.
20. Ni Ki C, Hosseinian-Far A, Daneshkhah A, Salari N. Topic modelling in precision medicine with its applications in personalized diabetes management. Expert Syst. 2022;39(4):e12774. doi: 10.1111/exsy.12774.
21. Li D, Li X. How to use LDA model to analyze patent information? Taking ships integrated power system as an example. In: Joshi A, Khosravy M, Gupta N, editors. Machine learning for predictive analysis. Springer; 2021. p. 51-64. doi: 10.1007/978-981-15-7106-0_6.
22. Govindarajan UH, Trappey AJ, Kumar G. Latent dirichlet allocation modeling for CPS patent topic discovery; 2019. p. 31-6. doi: 10.2991/icoiese-18.2019.6.
23. Wang J, Hsu CC. A topic-based patent analytics approach for exploring technological trends in smart manufacturing. J Manuf Technol Manag. 2020;32(1):Article 1. doi: 10.1108/JMTM-03-2020-0106.
24. Kim G, Park S, Jang DS. Technology analysis from patent data using latent dirichlet allocation. Adv Intell Syst Comput. 2014;271:71-80. doi: 10.1007/978-3-319-05527-5_8.
25. Yun J, Geum Y. Automated classification of patents: A topic modeling approach. Comput Ind Eng. 2020;147:106636. doi: 10.1016/j.cie.2020.106636.
26. Tian C, Zhang J, Liu D, Wang Q, Lin S. Technological topic analysis of standard-essential patents based on the improved Latent Dirichlet Allocation (LDA) model. Technol Anal Strateg Manag. 2022;36(9):2084-99. doi: 10.1080/09537325.2022.2130039.
27. Kumari R, Jeong JY, Lee BH, Choi KN, Choi K. Topic modelling and social network analysis of publications and patents in humanoid robot technology. J Inf Sci. 2021;47(5):Article 5. doi: 10.1177/0165551519887878.
28. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge University Press; 2008. doi: 10.1017/CBO9780511809071.
29. Dang B, Dang TT, Nguyen LM. SubTST: A combination of sub-word latent topics and sentence transformer for semantic similarity detection. In: Proceedings of the 14th international conference on agents and artificial intelligence. SCITEPRESS - Science and Technology Publications; 2022. p. 91-7. doi: 10.5220/0010775100003116.
30. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks; 2019. arXiv. arXiv:1908.10084. Available from: http://arxiv.org/abs/1908.10084.
31. Dang B, Le T, Nguyen LM. SubTST: A consolidation of sub-word latent topics and sentence transformer in semantic representation. Appl Intell. 2022;53(11):13470-87. doi: 10.1007/s10489-022-04184-x.
32. European Commission. Executive summary of the pharmaceutical sector inquiry report; 2009. [cited Mar 20 2023] Available from: http://ec.europa.eu/competition/sectors/pharmaceuticals/inquiry/communication_en.pdf.


**Cite this article:** Majhi D, Tiwari P, Mukherjee B. Measuring healthcare Innovations through Human Assigned Approach and Man-Machine Derived Approach: A Comparative Analysis Using Published Patents of India. J Scientometric Res. 2025;14(3):889-908.

**Annexure I: List of newly added Stopwords.**

| acceptable | brief | identification | moreover | quality | step |
|---|---|---|---|---|---|
| active | centre | immediate | nevertheless | real | subject |
| action | colour | improve | new | reduces | Suggest |
| agent | combination | include | packaging | related | sustained |
| analysis | comprise | increase | particularly | relates | system |
| apparatus | concerning | least | predominant | release | thereof |
| application | connect | little | preferably | sample | thereupon |
| apply | consist | low | preparation | say | unit |
| area | describe | mainly | prepare | seal | use |
| article | disclose | make | preparing | select | useful |
| assembly | facilitate | manifestation | process | sequence | user |
| base | free | mean | producing | show | whenever |
| basis | group | member | production | significant | whereafter |
| beginning | hereupon | method | provide | specific | wherein |
| breath | high | modify | pure | stable | whichever |

**Annexure II: Titles and Abstracts Topics.**

| Title Terms | Topic from Title fields | TN | Abstract Terms | Topic from Abstract fields |
|---|---|---|---|---|
| '0.111*"polymer" + 0.080*"food" + 0.058*"extend" + 0.052* "integrate" + ' '0.039*"derive" + 0.037*"recombinant" + 0.030* "light" + 0.027*"capsule" + 0.025*"vitamin" + 0.025*"dry" | Natural polymer to prepare capsule for drugs | 0 | '0.066*"hair"+0.031*"care"+0.021*"signal" + 0.018* "cleanse" + 0.015*"personal" + 0.015*"tissue" + 0.015* "cationic" + 0.014* "processing" + 0.014*"carbon" + 0.014* "region" | Cationic surfactants in hair conditioning formulation |
| '0.156*"anti" + 0.076*"activity" + 0.055*"target" + 0.046* "property" + 0.035*"colour" + 0.022*"bind" + 0.020* "inflammatory" + 0.020* "age" + 0.018*"flow" + 0.014* "plasma" | Preparation of anti-inflammatory pharmaceutical compositions/salts | 1 | '0.050*"water" + 0.038*"oil" + 0.031*"surfactant" + 0.030* "composition" + 0.028*"liquid" + 0.028*"phase" + 0.024* "aqueous" + 0.024*"product" + 0.022*"solution" + 0.021* "cosmetic" | Aqueous liquid cosmetic composition-grade surfactants |
| '0.097*"polymorphic" + 0.095*"health" + 0.062*"water" + 0.061* "purification"+ 0.043*"structure" + 0.023*"function" + 0.023*"surface" + 0.022*"fiber" + 0.022*"packaging" + 0.019*"improvement" | Methods for purification of amorphous water | 2 | '0.136*"alkyl" + 0.093*"substitute" + 0.065*"represent" + 0.043*"ring" + 0.040*"independently" + 0.038*"polymorphic" + 0.031*"aryl" + 0.026* "correspond" + 0.019*"unsaturated" + 0.016*"cycloalkyl" | Cycloalkyl Amine Compounds (Heterocyclic) for treatment of hyperproliferative diseases |
| '0.229*"skin" + 0.107*"composition" + 0.056*"emulsion" + 0.050* "surfactant" + 0.036*"produce" + 0.019*"solvate" + 0.019*"microbial" + 0.019*"specific" + 0.016*"prodrug" + 0.015*"sulfonamide" | Water-in-oil multiple emulsions and methods for use in adjuvants, pharmaceuticals, cosmetics, foods | 3 | '0.117*"device" + 0.071*"first" + 0.064*"user" + 0.061*"one" + '0.051* "configure" + 0.043*"plurality" + 0.033*"sensor" + 0.024*"element" + 0.024*"pressure" + 0.022*"parameter" | Sensors for sensing a plurality of parameters |
| '0.084*"peptide" + 0.081*"composition" + 0.073*"synergistic" + 0.043* "growth" + 0.039*"concentrate" + 0.034*"suspension" + 0.034* "induce" + 0.024*"anticancer" + 0.024*"nutritional" + 0.023* "ammonium" | Anticancer peptide drug development | 4 | '0.073*"composition" + 0.043*"extract" + 0.028*"relates" + 0.019* "combination" + 0.019*"plant" + 0.018*"formulation" + 0.015* "disclosure" + 0.014*"vaccine" + 0.014*"virus" + 0.012* "herbal" | Herbal extracted anti-viral agents/ Rotavirus vaccine composition |
| '0.262*"derivative" + 0.067*"management" + 0.058*"antibacterial" + 0.036*"portable" + 0.031*"prevention" + 0.023*"heart" + 0.022* "inhibition" + 0.018*"phosphate" + 0.015*"chronic" + 0.015* "manufacturing" | Preparation of nano-coating antimicrobial compound | 5 | '0.109*"weight" + 0.072*"layer" + 0.035*"surface" + 0.027* "substrate" + 0.024*"heart" + 0.022*"viscosity" + 0.022*"first" + 0.021*"molecular" + 0.021*"screen" + 0.021*"multi" | Cardiac tissue engineering/ Developing functional tissues for cardiac regeneration |

| Title Terms | Topic from Title fields | TN | Abstract Terms | Topic from Abstract fields |
|---|---|---|---|---|
| '0.384*"preparation" + 0.115*"composition" + 0.097*"product" + '0.056*"personal_care" + 0.038*"cell" + 0.035*"topical" + 0.032* "herbal" + 0.016*"extraction" + 0.016*"amine" + 0.015*"prevent" | Process of making herbal compositions | 6 | 0.045*"ch"+0.032*"hydrogen"+ 0.023*"bond" + 0.022*"cartridge" + 0.020*"colour" + 0.019*"branch" + 0.017*"unsubstituted" + 0.017* "acyl" + 0.016*"mask" + 0.016*"direction" | Hydrogen-bonding enhanced Nano-carrier drug delivery in vascular system, |
| '0.215*"hair" + 0.157*"composition" + 0.134*"salt" + 0.106*"care" + 0.070*"prepare" + 0.035*"synthesis" + 0.025*"tissue" + 0.018*"level" + 0.014*"inhibit" + 0.011*"obesity" | Method of preparing cosmetic compositions, therapeutic management of obesity | 7 | 0.053*"patient" + 0.049*"data" + 0.040*"body" + 0.032*"one" + '0.031*"medical" + 0.027*"device" + 0.025*"generate" + 0.024* "structure" + 0.021*"couple" + 0.018*"surface" | Hospital medical equipment/Devices |
| '0.290*"treatment" + 0.150*"composition" + 0.090*"disease" + 0.069*"cancer" + 0.048*"disorder" + 0.024*"treat" + 0.019*"human" + 0.017*"skin_lightene"0.016*"herbal" + 0.015*"blood"' | Herbal preparation of skin whitening compounds | 8 | '0.056*"protein" + 0.035*"blood" + 0.029*"cell" + 0.026*"level" + '0.021*"matrix" + 0.018*"polysaccharide" + 0.017*"therapeutic" + 0.014*"purification" + 0.013*"gene" + 0.013*"vitamin" | Protein based Therapeutics delivery agents |
| '0.380*"formulation" + 0.051*"patient" + 0.036*"pure" + 0.024* "oral" + '0.023*"prepare" + 0.018*"temperature" + 0.015*"medium" + 0.015*"indole" + 0.014*"animal" + 0.013*"suitable" | Pharmaceutically acceptable indole derivatives preparation and use as pesticides, fungicides, insecticides | 9 | '0.162*"composition" + 0.060*"pharmaceutical" + 0.050*"relates" + 0.047*"one" + 0.030*"polymer" + 0.026*"salt" + 0.024*"oral" + 0.020* "particle" + 0.019*"solid" + 0.018*"formulation" | Pharmaceutical preparation of solifenacin or a salt - Heterocyclic compounds |
| '0.117*"control" + 0.101*"oil" + 0.070*"hydrochloride" + 0.054* "monitor" + 0.045*"polymorph" + 0.035*"tablet" + 0.033* "antimicrobial" + 0.032*"body" + 0.027*"organic" + 0.024*"incorporate" | Polymorph drugs manufacture for treatment of infection | 10 | '0.064*"receive" + 0.056*"light" + 0.054*"chamber" + 0.036*"define" + 0.035*"tube" + 0.030*"source" + 0.028*"outer" + 0.024*"opening" + 0.023*"bar" + 0.020*"cavity" | Storing device for biological materials/ cavity resonator |
| '0.331*"comprise" + 0.220*"composition" + 0.083*"pharmaceutical" + 0.040*"enhance" + 0.023*"therapy" + 0.017*"sodium" + 0.011* "virus" + 0.011*"respiratory" + 0.011*"inflammation" + 0.010*"protection" | Method of treating/ preventing inhibiting respiratory and respiratory virus | 11 | '0.104*"acid" + 0.084*"preparation" + 0.055*"relates" + 0.052*"salt" + 0.034*"amino" + 0.034*"methyl" + 0.031*"intermediate" + 0.029* "amorphous" + 0.026*"yl" + 0.026*"formula" | Process for preparation of pharmaceutically accepted salts |
| '0.278*"form" + 0.108*"liquid" + 0.106*"preparation" + 0.077* "amorphous" + 0.074*"dosage" + 0.021*"injectable" + 0.021*"oral" + 0.018*"crystal" + 0.017*"calcium" + 0.015*"nutraceutical" | Formulation of oral dosages of amorphous compounds | 12 | '0.043*"two" + 0.038*"position" + 0.025*"lock" + 0.024*"perform" + '0.023*"mechanism" + 0.023*"cover" + 0.020*"load" + 0.019*"open" + 0.019*"location" + 0.018*"catheter" | Device based on Lock and release mechanisms for trans-catheter implantable devices/ Surgical instruments |
| '0.329*"device" + 0.099*"treat" + 0.070*"extract" + 0.051*"medical" + 0.048*"condition" + 0.044*"manufacture" + 0.034*"plant" + 0.025*"amino" + 0.021*"film" + 0.011*"diabetes" | Plant extract methods and use for treatment of diabetics and other diseases | 13 | '0.035*"treatment" + 0.028*"cancer" + 0.027*"disease" + 0.021* "cell" + 0.019*"relates" + 0.012*"condition" + 0.011*"inflammatory" + 0.010* "anti" + 0.010*"human" + 0.010*"treat" | Anti-Inflammatory Agents for Cancer Therapy |
| '0.157*"solid"+0.123*"high"+0.102*"co njugate"+ 0.032* "spray"+0.023*"taste_ maske"+0.022*"antiviral"+ 0.021* "enhancement" + 0.019*"alpha" + 0.018*"antidiabetic" + 0.015*"powder"'), | Preparation of plant based anti-diabetic, antiviral nono-formulations | 14 | '0.112*"drug" + 0.065*"delivery" + 0.039*"target" + 0.038* "conjugate" + 0.030*"dispersion" + 0.030*"injection" + 0.026* "healthcare" + 0.026* "polypeptide" + 0.017*"droplet" + 0.017* "fragment" | Components for targeted drug delivery system |

| Title Terms | Topic from Title fields | TN | Abstract Terms | Topic from Abstract fields |
|---|---|---|---|---|
| '0.446*"composition" + 0.172*"pharmaceutical" + 0.081*"acid" + 0.073*"oral" + 0.028*"particle" + 0.023*"crystalline" + 0.011*"cream" + 0.007*"ester" + 0.007*"sugar" + 0.005*"transfer" | Crystalline pharmaceutical composition of drugs, creams etc. | 15 | '0.130*"compound"+0.092*"formula"+0.061*"composition" + 0.046* "relates" + 0.044*"pharmaceutical" + 0.039*"salt" + 0.036* "treatment" + 0.032*"disease" + 0.028*"disorder" + 0.023* "derivative" | Pharmaceutical salts of pyrimidine derivatives and method of treating disorders |
| '0.119*"intermediate"+0.112*"preparation "+ 0.072*"therapeutic " + 0.066*"aqueous" + 0.048*"stabilize" + 0.039*"gel" + 0.038* "nanoparticle" + 0.033*"metal" + 0.031*"coat" + 0.025*"kit" | Composition and use of surface modified meta nanoparticples | 16 | '0.073*"wt" + 0.036*"bacterial" + 0.028*"anti" + 0.028*"treat" + "0.023*"fiber" + 0.023*"handle" + 0.022*"continuous" + 0.020* "resistance" + 0.020*"resistant" + 0.019*"bacteria" | Nanocomposite inhibitors for antibacterial', 'antifungal', activity |
| '0.168*"drug" + 0.165*"delivery" + 0.084*"receptor" + 0.037* "analog" + 0.025*"bone"+0.022*"resistant" + 0.018*"pathogen" + 0.013* "regulate" + 0.012*"excipient" + 0.012*"healthy" | Targeted drug delivery methods and devices | 17 | '0.081*"crystalline" + 0.046*"peptide" + 0.044*"high" + 0.039* "relates" + 0.037*"preparation" + 0.028*"yield" + 0.024*"purity" + 0.019*"hiv" + 0.018*"simple" + 0.018*"cost" | Preparation of therapeutic peptide for clinical purpose |
| '0.106*"production" + 0.074*"complex" + 0.061*"material" + 0.040* "mechanism"+ 0.034*"ingredient" + 0.033*"biological" + 0.029* "mediate" + 0.022*"probiotic" + 0.019*"vector" + 0.019* "oxygen" | Molecular complex preparation and use | 18 | '0.033*"silicone" + 0.028*"determine" + 0.020*"medicament" + 0.015* "intravenous" + 0.015*"surgical" + 0.014*"transdermal" + 0.013*"valve" + 0.013*"shampoo" + 0.012*"range" + 0.011*"implant" | Silicon-based devices for delivery of therapeutic agents, Herbal preparation for cosmetics. |
| '0.251*"compound" +0.159*"inhibitor"+0.076*"derivative"+ 0.075*"substitute"+0.043*"conditioning"+ 0.042*"protein"+ 0.039*"modulator"+0.019*"fuse"+ 0.019*"kinase" + 0.019*"low" | Protein kinase inhibitor | 19 | '0.164*"skin" + 0.086*"image" + 0.081*"acid" + 0.061*"fatty" + 0.027* "material" + 0.025*"anionic" + 0.023*"antibody" + 0.020*"cream" + 0.016*"sodium" + 0.013*"wound" | Medical preparation of sodium alginate for wound healing |

TN=Topic Number.