Analysis of Clusters With Indian Patent Data Using Different Word Embedding Techniques

Received: 4 Sep, 2023 **Revised: 12 Nov, 2023** **Accepted: 1 Dec, 2023**

Pankaj Beldar[1*], Mohansingh Pardeshi[2], Rahul Rakhade[3], Shilpa Mene[4]

[1*]K.K.Wagh Institute of Engineering Education and Research -
Email:- prbeldar@kkwagh.edu.in
[2]K.K.Wagh Institute of Engineering Education and Research
Email:- mrpardeshi@kkwagh.edu.in
[3]K.K.Wagh Institute of Engineering Education and Research
Email:- rdrakhade@kkwagh.edu.in
[4]K. K. Wagh Institute of Engineering Education & Research,Nashik
Email:- spmene@kkwagh.edu.in

***Corresponding Author:** Pankaj Beldar
*K.K.Wagh Institute of Engineering Education and Research
prbeldar@kkwagh.edu.in

Abstract.
This study employs advanced Unsupervised Machine Learning (UML) techniques, including K-means and Agglomerative clustering, to analyze descriptive Indian Patent data. Utilizing silhouette score evaluation, elbow method, and dendrogram analysis, optimal cluster numbers are determined. Various word embedding methods like TF-IDF, Word2Vec, and Countvectorizer, combined with rigorous text processing, are explored. Robust testing of categorical and numerical features yields a high silhouette score of 0.8965 for 2 clusters, showcasing Agglomerative clustering's effectiveness. The research emphasizes the crucial role of UML techniques, word embedding methodologies, and comprehensive text processing in revealing complex structures within Indian Patent data. Besides advancing unsupervised learning methodologies, this work aids scholars, practitioners, and policymakers in comprehending the Indian patent landscape, fostering innovation, and technological progress.

Keywords: K-means, Agglomerative clustering, Word embedding, Patents, Silhouette Score

1 Introduction

In a time of invention and rapid technological development, intellectual property rights (IPR) and patents are essential for preserving the arts, promoting development, and propelling global economic expansion. In the complex web of international economies, India stands out as a country where the significance and effect of intellectual property rights, especially patents, have changed dramatically. In the socioeconomic environment of India, the value of intellectual property rights which include patents, copyrights, trademarks, and trade secrets—cannot be emphasized. In recent decades, India has experienced a growing focus on innovation in a variety of fields, which has raised awareness of the need of protecting and utilizing intellectual property. The objective of this study paper is to examine the various aspects of intellectual property rights (IPR) and patents in Indian society, clarifying their significance, ramifications, and the changing dynamics within the nation's socio-economic structure. In order to decipher the complex web of effects and outcomes surrounding intellectual property rights in India, this study will look at the subtle interactions that exist between IPR laws, innovation, entrepreneurship, and national development. The study will examine the legislative framework, policy interventions, and their effects on innovation and economic growth as it moves through the historical development of India's intellectual property landscape. This article will also include case studies and actual data to highlight the concrete effects of IPR and patents on a variety of industries, including the Indian creative industries, technology, pharmaceuticals, and agriculture. The use of natural language processing (NLP) makes it easier to extract important data from policy frameworks, patent databases, and legal papers. This helps identify new technologies, innovation patterns, and how Indian IPR laws are changing. In order to better comprehend the implications and interpretations of IPR laws and patent applications in the Indian socio-economic context, researchers can do sentiment analysis, entity recognition, summarization, and categorization of legal texts by using NLP-powered algorithms. Furthermore, by utilizing word embedding methods like Word2Vec, GloVe, or BERT, words and phrases can be represented as high-dimensional vectors that capture contextual meanings and semantic relationships found in the textual corpus. With the use of these methods, scholars can investigate the semantic parallels, groups, and connections found in legal and patent texts, enabling a more sophisticated examination of the terminology found in Indian IPR frameworks, court rulings, and patent

specifications. The application of natural language processing (NLP) and word embedding methods to the review of Indian patents and intellectual property rights complements conventional analysis by offering computer resources that facilitate the extraction of meaningful patterns, correlations, and insights more quickly. The utilization of an interdisciplinary method not only improves research efficiency but also broadens our understanding of the intricate relationship between language, regulations, and innovation in the context of intellectual property in Indian culture.

2 Literature Review

In the heart of India's vibrant innovation ecosystem, the year 2023 witnessed a remarkable surge in inventive brilliance, as reflected in the multitude of patents granted throughout the year. The Government of India, recognizing the pivotal role of patents in fostering technological advancements and economic growth, diligently compiled and made available a comprehensive dataset detailing the weekly patent applications granted during this transformative period. (Grander et al., 2021) had focused on dataset, a testament to the nation's creative ingenuity, provides a unique window into the dynamic landscape of nations innovation. This research paper embarks on an insightful exploration of the dataset sourced from data.gov.in, focusing on the weekly patent applications granted in India throughout 2023. Each entry within this dataset represents a milestone, encapsulating not just legal proceedings but the culmination of inventive ideas, scientific discoveries, and entrepreneurial ambitions. With this dataset as our guide, our study seeks to unravel hidden patterns, discern trends, and extract invaluable insights that illuminate the multifaceted aspects of India's innovation panorama. In the context of this dataset, our research endeavours to uncover the intricate stories woven into the fabric of each patent application. As we delve deep into the data, we aim to identify thematic clusters, pinpoint dominant technological domains, and explore the geographical distribution of innovation across different regions of India .Through careful study, we will uncover temporal trends that will shed light on how technology will develop over the course of 2023 and give us a dynamic view of the country's creative journey. This investigation pays homage to India's innovative spirit and goes beyond a simple statistics investigation. It honours the innovators, researchers, and businesspeople whose unrelenting quest for excellence drives the development of the country. This study aims to deepen our knowledge of India's innovative landscape by analysing the weekly patent applications issued in 2023. It aspires to offer useful information for researchers, policymakers, entrepreneurs, and innovators alike, empowering them to make wise choices, promote collaborations, and actively contribute to the country's innovation-driven growth. Our goal in writing this research paper is to advance the conversation about Indian innovation by highlighting its variety, resiliency, and forward-thinking outlook. (Feng, 2023)By examining the weekly patent granted, we analyse the past while also laying the foundation for a future in which knowledge, creativity, and collaboration meet to drive India's innovation agenda and create a better future for future generations. Subsequent paragraphs, however, are indented.(Lone & Warale, 2022) had examined K-means and agglomerative clustering techniques that have been shown to be essential tools for pattern identification and data analysis in recent studies. Large datasets can be efficiently clustered using the centroid-based K-means clustering technique, which divides data into discrete groups according to similarities. However, a hierarchical technique called agglomerative clustering creates larger clusters by combining smaller ones, providing information on complex linkages found in the data. Studies comparing various approaches have demonstrated their distinct advantages and uses, assisting researchers in selecting the best approach for their datasets. These studies are essential for determining the future of data-driven decision-making in a variety of sectors, from market segmentation in business to comprehending biological data(Cahyo & Sudarmana, 2022) had used these unsupervised techniques for Comparison of K-Means and Agglomerative Clustering for Users Segmentation based on Question Answerer Reputation in Brainily Platform.

Therefore, examining Indian patents is essential for understanding the country's innovation ecosystem, policy implications, legal frameworks, and societal impact, shaping discussions and decisions at national and international levels.

3 Methodology

3.1 Dataset

One of the primary platforms for accessing free government data in India is the Open Government Data (OGD) Platform (data.gov.in). This platform is managed by the National Informatics Centre (NIC) and provides access to a wide range of datasets from various government departments and ministries. Dataset is taken from the website (Shri Rajesh Kumar Sharma, 2023) https://data.gov.in/resource/weekly-patent-application-granted-during-2023 . This dataset updates on weekly basis. We have taken data up to September 2023 for analysis. On the data.gov.in website, datasets related to diverse topics such as agriculture, health, education, finance, environment, transportation, and many others. The datasets are available in different formats and can be downloaded for analysis, research, and other purposes. The patent Dataset has attributes as -The Data Frame contains 35,206 entries (rows) and 11 columns. Columns consist of various data types: object (strings), datetime64 (date and time), and float64 (floating-point numbers). Features are as follows-

'PUBLICATION_NUMBER': This column likely contains unique identification numbers assigned to each published patent document. It serves as a unique identifier for patents. 'PUBLICATION_DATE': This column indicates the date when the patent application was published and made publicly available. It's an essential timestamp for tracking the publication timeline of patents. 'IPO_LOCATION': IPO stands for Intellectual Property Office. This column specifies the geographical location or jurisdiction of the Intellectual Property Office where the patent application was filed or granted. It indicates the country or region's patent office. 'APPLICATION_TYPE_DESC': This column provides a description of the type of patent application. It could indicate whether the application is a national application, an international PCT (Patent Cooperation Treaty) application, or any other specific application type.

'APPLICATION_NUMBER': An individual number assigned to every patent application is contained in this column. It is used internally by the patent office to track the application throughout the patenting process and is distinct from the publication number. "DATE_OF_FILING": The date when the patent application was formally submitted to the Intellectual Property Office is indicated in this column. The process of applying for a patent officially begins at this point. "TITLE_OF_INVENTION": The name or title of the invention that is mentioned in the patent application is contained in this column. It gives a succinct overview of the novel idea or technological advancement that is protected by the patent. "FIELD_OF_INVENTION": This column identifies the area of technology or field to which the invention belongs. It assigns the patent to a certain field, such electronics, chemistry, mechanical engineering, or biotechnology. 'NO_OF_PAGES': The total number of pages of the patent document is shown in this column. It is an indicator of the complexity and length of the patent application, frequently illustrative of the extensive technical description and claims. 'NO_OF_CLAIMS': The total number of claims included in the patent application is listed in this column. Patent claims specify the boundaries of the security provided by the patent as well as the features of the invention that are shielded from infringement. 'DATE_UPDATED_d_m_y': This column likely contains the date when the patent information or status was last updated. It's essential for tracking the most recent changes or developments related to the patent application. We have used the methodology shown in the table 1 for initial analysis because there has not been any studies done on this dataset. Every exploratory data finding is first tested.

Table 1. Methodology

Methodology	
Data Pre-Processing	Data Pre-processing includes- Handling of Missing Values Handling Outliers Removing Duplicate Values Feature Extraction from Date and Time
Exploratory Data Analysis	Finding of Field wise Patent Count Finding Correlation between number of pages and number of claims Finding IPO Location wise patent count Finding Patent Application wise Patent count Finding of Field wise Days required for publication
(NLP)Natural Language Processing(Volodymyr Zhukov, 2023)	1. Lowercasing: Change the text's case. Split text into words or tokens by using tokenization. 3. Eliminating Symbols and Numerical Digits: Remove special characters and numbers. 4. Eliminating Stop words: Get rid of words like "and," "the," and "was." 5. Stemming: Convert words to their simplest form, for as "running" to "run." 6. Lemmatization: It is the process of shortening words to their dictionary equivalents, such as "running" to "run." 7. Expand contractions (such as "I'll" to "I will") when handling them. 8. Replace with placeholders ('URL', 'EMAIL') when handling URLs and email addresses. 9. Remove HTML mark-up by removing HTML tags. 10. Eliminating or Correcting Typos: Eliminate non-dictionary words or fix frequent typos. 11. Standardize the use of whitespace by eliminating superfluous spaces. 12. The CountVectorizer for Feature Extraction 13. Text to numerical vector conversion (Word2Vec, TF-IDF).
Word Embedding Techniques (Shubham Chouksey, 2020; Volodymyr Zhukov, 2023)	Bag of Words TF-IDF (Term Frequency-Inverse Document Frequency) Word2Vec
Clustering Analysis	K-means algorithm Agglomerative Clustering
Results	
Discussion	

4 Exploratory data Analysis

(Munshi et al., 2022)Exploratory Data Analysis (EDA) is essential for comprehending and interpreting datasets pertaining to patents. EDA helps identify patterns and trends within patent data. It allows researchers to discern recurring themes, technological shifts, and emerging areas of innovation, providing valuable insights for inventors, businesses, and policymakers. EDA helps assess the quality of the patent data. By visualizing the distribution of patents across categories, regions, or time, researchers can identify outliers, inconsistencies, or missing data points, ensuring the dataset's reliability.(Sahoo et al., 2019) For machine learning applications, EDA aids in selecting relevant features. By analyzing correlations and relationships between different patent attributes, researchers can choose the most influential features for predictive models, enhancing the accuracy of patent-related predictions.(Da Poian et al., 2023) EDA enables the visualization of patent distribution across different geographical regions. This geographical analysis is vital for businesses and policymakers, helping them identify innovation hubs, plan investments, and formulate regional development policies. The examination of patent patterns over time is made easier by EDA. By visualizing the number of patents filed or

granted annually, quarterly, or monthly, researchers can identify temporal patterns, technological cycles, and the impact of policy changes on patent activities. Understanding these columns is essential for conducting in-depth analyses, such as identifying patent trends, assessing the scope of inventions, or conducting geographical and technological analyses within the field of patents.

4.1 Handling Missing Values

(Sahoo et al., 2019) When some details are missing in the data, we tidy it up before studying it. For specific columns like the number of pages or claims, and the title or field of invention, we replace the missing info with other values. For example, if we don't know the number of pages or claims, we use the average numbers for those. If the title of an invention is missing, we write 'unknown' to show we don't have that info. And if the type of invention is missing, we put the most common type, 'MECHANICAL', to keep things consistent. This helps make sure the data is complete and ready for analysis.

4.2 Handling Duplicate values

Dataset has 20 duplicate values hence for further analysis duplicate values are removed.

4.3 Handling outliers

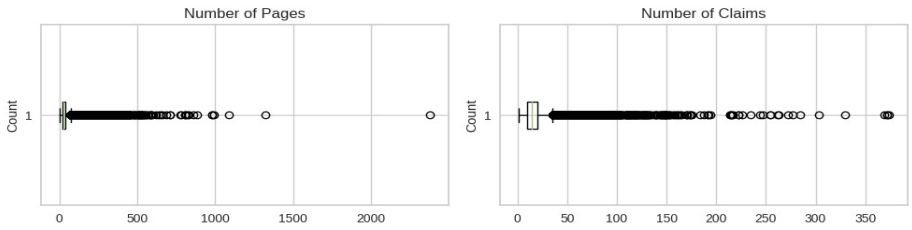

Fig. 1. Number of Pages and Claim Data in a Boxplot

Fig. 1 displays a boxplot of the number of pages and claims; the majority of the data appears to be an outlier, but because it is real-time data from a government database, we decided to keep it the same for further study.

4.4 Field wise Patent distribution

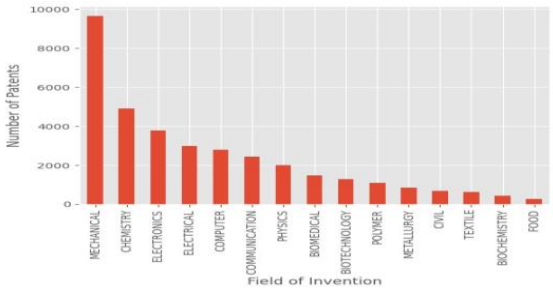

Fig. 2. Field wise Patent distribution

In India, patents cover various fields like Mechanical, Chemistry, Electronics, Electrical, Computer, and more. The most patents are in Mechanical, Electronics, Electrical, and Computer fields, showing a focus on technology and engineering. Some fields like Biomedical and Biochemistry highlight ongoing healthcare research. Fewer patents in fields like Food and Textile suggest areas for potential growth. Patents in Communication and Electronics reflect their importance in technology and the economy, while Physics and Metallurgy show ongoing scientific research. Overall, different fields indicate diverse innovation areas, but more detailed analysis would give a clearer picture.

4.5 IPO Location wise Patent Distribution

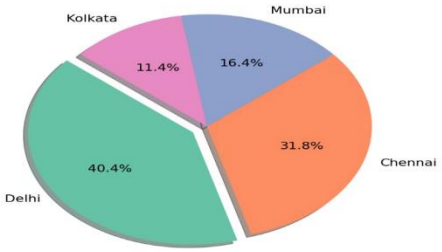

Fig. 3. IPO Location wise Patent Distribution

Delhi leads with the most patents filed, showing strong innovation. Chennai follows closely, indicating active research. Mumbai and Kolkata have fewer patents but still show consistent innovation. Different cities specialize in different areas of research. Good policies in some places attract more innovation. Cities with fewer patents have room to grow by encouraging more research.

4.6 Means days required to publish by Field of Invention

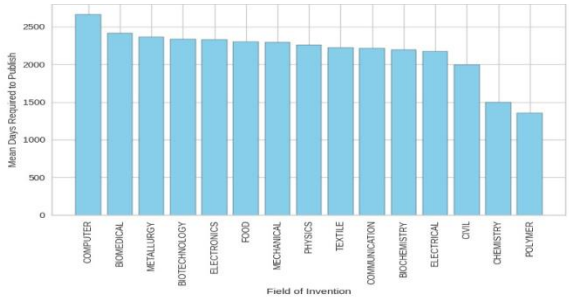

Fig. 4. Means days required to publish by Field of Invention

In Fig. 4, we see how long it takes for inventions in different fields to get published. Polymer research is the quickest, taking around 1355 days. This could be because it's focused and easier to review.Computer science takes the longest at about 2662 days. It's complex and changes quickly, needing thorough testing and review before new ideas are accepted.Chemistry, Civil Engineering, and Electrical Engineering also take a while, about 1495 to 1991 days. This might be because they need detailed testing and review due to their complex nature.On the other hand, fields like Polymer and Biochemistry have shorter times. They might have simpler review processes or research that's easier to check.This data shows how different fields need different amounts of time for their inventions to get published. It highlights how each field has its own complexities and methods that affect how long it takes for new ideas to become public.

4.7 Count of Different Application Types

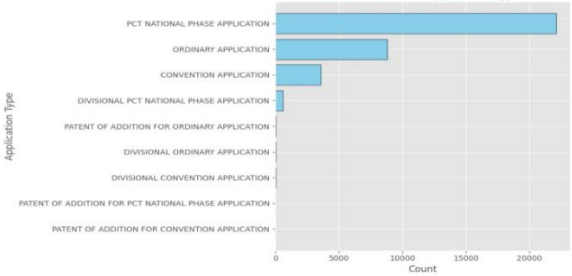

Fig. 5. Count of Different Application Types

From fig. 5 it is evident that the majority of patent applications fall under the category of PCT NATIONAL PHASE APPLICATION, indicating a substantial international interest in protecting inventions across different countries. ORDINARY APPLICATIONS also represent a significant portion, demonstrating a substantial number of inventions intended for domestic protection. CONVENTION APPLICATIONS are also notable, although to a lesser extent, indicating a focus on inventions protected within specific treaty-conforming countries. The presence of DIVISIONAL APPLICATIONS suggests instances where initial patent applications have been divided into multiple applications, likely due to the complexity or diversity of the invention. Additionally, the small counts in categories like PATENT OF ADDITION FOR ORDINARY APPLICATION, DIVISIONAL ORDINARY APPLICATION, DIVISIONAL CONVENTION APPLICATION, PATENT OF ADDITION FOR PCT NATIONAL PHASE APPLICATION, and PATENT OF ADDITION FOR CONVENTION APPLICATION indicate relatively specialized cases where modifications or additions are made to existing patent applications, showcasing the diverse nature of patent-related activities.

4.8 Exploring of Patent Titles

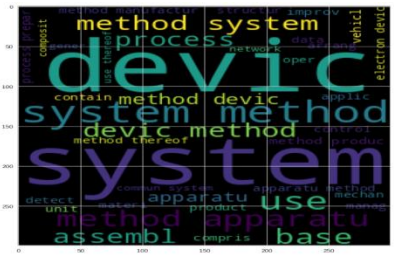

Fig. 6. Word cloud of Patent Titles

Table 2. Most frequent words in Patent Title

Sr. No.	Word	Count
1	method	12426
2	system	6824
3	device	6475
4	apparatus	2758
5	process	2400
6	use	2311
7	control	2308
8	composite	1830
9	thereof	1763
10	vehicle	1654

5 Word embedding and Clustering Analysis

5.1 Word2Vec

(Tezgider et al., 2019)A common word embedding method in natural language processing (NLP) is called Word2Vec, which expresses words as dense, high-dimensional vectors of real numbers. For numerous NLP tasks, including text categorization, language translation, and sentiment analysis, these vectors record the semantic associations between words.

Tezgider et al., 2019)The foundation of Word2Vec is the idea that words with related meanings frequently appear in settings. Words are modeled as points in a high-dimensional space using the concept of distributed representations. The words here are grouped together by similarity of meaning.

(Gagliardi & Artese, 2020; Mohammadi et al., 2021)Two primary algorithms make up Word2Vec: Continuous Bag of Words (CBOW) and Skip-gram. Using the context words (words that come before and after the target word), CBOW predicts the target word. It gains the ability to guess a target word depending on the words around it. On the other hand, based on the target word, skip-gram guesses the words in the surrounding sentences. When a target word is provided, it learns to anticipate the context terms. In situations involving huge datasets, skip-gram is frequently preferred.

(Wahba et al., 2020)CBOW and Skip-gram are two examples of shallow neural networks with only one hidden layer. One-hot encoded vectors that represent words make up the input layer. The target or context words are predicted by the output layer, while the hidden layer holds the word vectors that need to be learned.

(Tezgider et al., 2019)Word2Vec's learning goal is to learn word vectors that minimize the negative log likelihood of the target word given the observed context words (or the opposite for Skip-gram). The word vectors are iteratively adjusted during the training process to ensure that similar words have similar vector representations.

(Kumar et al., 2018; Onan & Toçoğlu, 2021; Tache et al., 2021)Semantic similarity computations are made possible by the similar vector representations of words with like meanings. For instance, the semantic similarity between words can be measured by the cosine similarity between word vectors.

(Kumar et al., 2018)Various linguistic associations, including synonyms, antonyms, and analogies, are captured in the high-dimensional space in which words are embedded. The proximity of words with similar meanings in this space is reflected by vector operations, which also reflect linguistic links. Word2Vec models are effective tools for applications requiring natural language understanding because they can capture semantic nuances and can be trained on big textual datasets. By enabling algorithms to interpret language in a way that captures semantic meanings and relationships, these learnt word embedding have revolutionized the area of NLP and enhanced performance in a variety of applications.

5.2 TF-IDF (Term Frequency-Inverse Document Frequency)

(Kalra et al., 2022)It is a metric that measures the weighting of a word in a document in relation to a corpus of documents or a particular document within a corpus.

Term Frequency (TF): (Kalra et al., 2022)This measures how frequently a term occurs in a document. It is calculated as the number of times a term t appears in a document d divided by the total number of terms in the document. It can be represented as:(Shubham Chouksey, 2020)

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ Number\ of\ terms\ in\ document\ d}$$

Inverse Document Frequency (IDF):
(Kalra et al., 2022)This measures how important a term is by considering how often it appears across multiple documents in a corpus. If a term appears in many documents, it is considered less important. IDF is calculated as the logarithm of the total number of documents (N) divided by the number of documents containing the term (n_t). It can be represented as:(Shubham Chouksey, 2020)

$$IDF(t,D) = \log\left(\frac{N}{n_t}\right)$$

Where *N* is the total number of documents in the corpus and n_t is the number of documents containing the term *t*.

TF-IDF Score:
(Kalra et al., 2022) The TF-IDF score of a term in a document is calculated by multiplying its TF and IDF scores:(Shubham Chouksey, 2020)

$$TF - IDF(t,d,D) = TF(t,d) * IDF(t,D)$$

A term's importance to the document is shown by its TF-IDF score, which increases as a term's score increases. For tasks like text categorization, clustering, and document relevance ranking in search engines, TF-IDF is frequently employed in information retrieval and text mining. It aids in locating significant terms within a corpus of documents and is especially helpful when basic word frequency metrics are insufficient or inaccurate.

5.3 Text processing:
(S. Li & Gong, 2021; Volodymyr Zhukov, 2023)Natural Language Processing (NLP) uses text processing as a vital preprocessing step to convert raw text input into a format appropriate for analysis and machine learning applications. This process uses a number of tools and procedures intended to clean, tokenize, normalize, and vectored textual data.
The initial stage of text processing is tokenization, which involves breaking down a raw text into tokens, or discrete words, phrases, or sentences. By dividing the text into meaningful tokens, tokenization makes it possible to do more in-depth analysis.

Normalization: By making all characters lowercase, eliminating punctuation, and extending contractions, normalization tries to assure consistency in the text data. Normalization improves the accuracy of subsequent studies by standardizing the text, which lowers the complexity of the data and assures that related terms are processed in the same way.
Stop word removal: Common words with limited semantic value, such as "and," "or," and "the," are frequently eliminated during text processing. Stop words are eliminated to reduce noise in the data and concentrate analysis on important words and phrases.

Lemmatization and stemming are methods for breaking down words into their component parts, respectively. Lemmatization reduces words to their dictionary or base form, whereas stemming includes eliminating prefixes and suffixes from words. These procedures make words uniform, guaranteeing that various inflected forms are rendered consistently.
Vectorization: The text must be transformed into numerical vectors that machine learning algorithms can interpret once it has been tokenized and normalized. Text can be transformed into numerical features using methods such as Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). BoW depicts text as a matrix of word occurrences, whereas TF-IDF emphasizes uncommon and significant terms while taking into account the significance of words over the entire corpus.

Named Entity Recognition (NER) is a particular text processing technique that finds and groups named entities in text, such as names of people, organizations, and locations. NER is crucial for extracting specific information from unstructured text input and is therefore necessary for applications like information retrieval and question-answering systems.
In order to analyze, model, and interpret unstructured text input, text processing in NLP is a multi-step procedure that transforms information into a structured and numerical format. These foundational methods are used by a wide range of applications, such as sentiment analysis, chatbots, language translation, and information retrieval systems, to draw conclusions from vast amounts of textual data. Proper text processing enhances the precision and effectiveness of NLP algorithms, enabling a deeper comprehension of human language in the digital realm.

5.4 One hot encoding
(Mehta et al., 2021)In order for machine learning algorithms to handle categorical variables correctly, one-hot encoding, a fundamental data preparation technique, is essential. The non-numeric labels that make up categorical variables, such "FIELD_OF_INVENTION" and "IPO_LOCATION," denote several categories or classes. One-hot encoding is useful

in situations when machine learning algorithms need numerical input. By converting categorical variables into a binary format using this technique, each distinct category is represented as a binary vector. One-hot encoding, for instance, would produce distinct binary columns for each category if a column like "FIELD_OF_INVENTION" contained categories like "COMPUTER," "BIOMEDICAL," and "METALLURGY." Then, each row in the dataset is represented by a binary vector, with a "1" denoting the existence of a certain category and "0s" in all other locations.

5.5 Principle Component Analysis

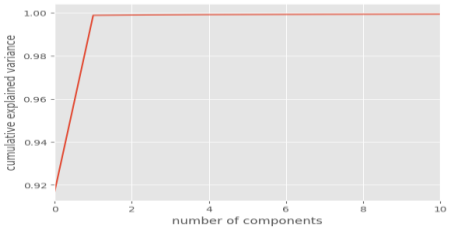

Fig. 7. Principle Component Analysis

PCA is a tool that simplifies complex data by finding important patterns. It helps reduce data size while keeping key details intact. This makes it easier to understand and analyze large datasets in fields like genetics, economics, and image processing.

5.6 Clustering

(Mehta et al., 2021)A measure of a clustering algorithm's quality is its silhouette score. It sheds light on how far apart the resulting clusters are from one another. A high value means the object is well matched to its own cluster and poorly matched to nearby clusters. The silhouette score goes from -1 to +1. The silhouette score is determined as follows:(Imad Dabbura, 2018)
Here is how to calculate the silhouette score for a single data point, i:

$$\text{Silhouette score (i)} = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

The average distance a(i) between the i-th data point and the other data points in the same cluster. When minimizing over clusters, b(i) is the lowest average distance between the i-th data point and the data points in the other cluster.

Silhouette score

(Mu et al., 2022) The average of the silhouette scores for each data point makes up the overall silhouette score. If the score is close to +1, the sample is remote from the nearby clusters. A score of 0 denotes that the sample is either on or very near the decision boundary between two adjacent clusters. A result that is nearly -1 suggests that the samples may have been placed in the incorrect cluster. Higher silhouette ratings typically suggest clusters that are better delineated. It is significant to note that the silhouette score can be useful in determining the ideal number of clusters for a dataset even when the true number of clusters is unknown. It does have certain restrictions though, particularly when working with irregularly sized clusters or non-convex structures. In order to completely assess the quality of clustering results, it is frequently used in conjunction with other metrics and visualizations.

5.7 K-means Clustering

(M. J. Li et al., 2008)Iterative algorithms like K-means clustering are frequently employed in unsupervised data analysis and pattern detection. It works by dividing a dataset into K different clusters, each of which is represented by a feature space point called the centroid. The process starts by picking K initial centroids at random. Then, using the Euclidean distance formula, it determines the distance between each data point and each centroid:(Imad Dabbura, 2018)

$$Distance(x, c_i) = \sqrt{\sum_{j=1}^{n} \left(x_j - c_{ij}\right)^2}$$

Where x is a data point, c_i denotes the feature's centroid at it, x_j denotes its jth feature, and n denotes the total number of features. The closest centroid is given to each data point. The centroids are updated by calculating the mean of all the points assigned to each cluster after they have all been assigned. Until the centroids converge, which means they no

longer change appreciably, or a maximum number of iterations is reached, the assignment and centroid updating procedure iterates. The sum of squared distances between data points and their respective centroids, also known as the inertia or within-cluster sum of squares is the goal function being minimized during this process:(Imad Dabbura, 2018)

$$Inertia = \sum_{i=1}^{K} \sum_{x \epsilon C_i} ||x - c_i||^2$$

Where c_i represents the set of data points assigned to the ith cluster and $||x - c_i||^2$ denotes the squared Euclidean distance between a data point x and its assigned centroid c_i. K-means is widely used in various fields for tasks like customer segmentation, image compression, and anomaly detection, providing valuable insights into the underlying patterns of the data.

5.8 Agglomerative clustering

(Cahyo & Sudarmana, 2022)A popular hierarchical clustering technique in machine learning and data analysis is agglomerative clustering. Agglomerative clustering creates a hierarchy of clusters, unlike partitioning techniques like K-means. Starting with treating each data point as a separate cluster, it iteratively merges the closest clusters using a similarity metric until a single cluster that includes all the data points is created. A dendrogram, a tree-like diagram that depicts the clustering of data, can be used to visualize the process.

(Reddy et al., 2017)Calculating the distances between clusters is at the heart of agglomerative clustering. The distance between clusters can be calculated using a variety of linkage criteria. Complete linkage is a popular strategy in which the distance between two clusters A and B is the greatest distance possible between any two places, one from each cluster:

Distance (A, B) =max (distance (a,b)) for a∈A,b∈B

Another approach is single linkage, where the distance between two clusters A and B is the minimum distance between any pair of points, one from each cluster:

Distance (A, B) =min (distance (a,b)) for a∈A,b∈B

The average linkage criterion calculates the distance between two clusters A and B as the average distance between all pairs of points, one from each cluster:

$$\text{Distance (A, B)} = \frac{1}{|A| \times |B|} \sum_{a \in A} \sum_{b \in B} \text{distance}(a, b)$$

Here, ||A|| and ||B|| represent the number of points in clusters A and B respectively. The choice of linkage criterion significantly influences the shape and characteristics of the clusters. Once the distances are calculated, clusters are merged iteratively until a stopping criterion, often a desired number of clusters or a specific distance threshold, is reached.

6 Results

6.1 Clustering with CountVectoriser and K-Means

The outcomes of K-means clustering studies using Count Vectorizer applied to various feature sets. As evident, the silhouette score, a metric indicating cluster cohesion and separation, varies based on the features considered. Notably, the inclusion of specific attributes, such as 'No. of pages' and 'No. of Claims,' significantly improves clustering quality, as indicated by the high silhouette score of 0.8728 achieved with four features ('Title of the Invention, IPO Location, No. of pages, No. of Claims, Field of Invention') resulting in two optimum clusters. This outcome underscores the importance of careful feature selection, with a focus on meaningful attributes, in enhancing the efficacy of clustering algorithms. Further analysis and inference based on these results could provide valuable insights into the patent data, aiding researchers and practitioners in understanding patterns and relationships within inventions, IPO locations, and related attributes. Clusters between 2 and 10 are selected for all analyses because, according to the results, clusters larger than 10 have a very low silhouette score.

Table 3. Silhouette score results with Count Vectoriser and diffefent set of Features

Cluster Number	Title of the invention	Title of the Invention, IPO Location	Title of the Invention, IPO Location, Field of Invention	Title of the Invention, IPO Location, No. of pages, No. of Claims, Field of Invention
Cluster 2	0.06761	0.04748	0.04215	0.87287
Cluster 3	0.05074	0.04214	0.03634	0.71194
Cluster 4	0.04865	0.07241	0.05005	0.58385
Cluster 5	0.04838	0.04806	0.03355	0.53497
Cluster 6	0.02710	0.04737	0.03247	0.46429
Cluster 7	0.02589	0.05147	0.04118	0.44766
Cluster 8	0.04022	0.05215	0.02308	0.44595

Cluster 9	0.03093	0.04423	0.03170	0.38338
Cluster 10	0.03144	0.05050	0.02102	0.38550

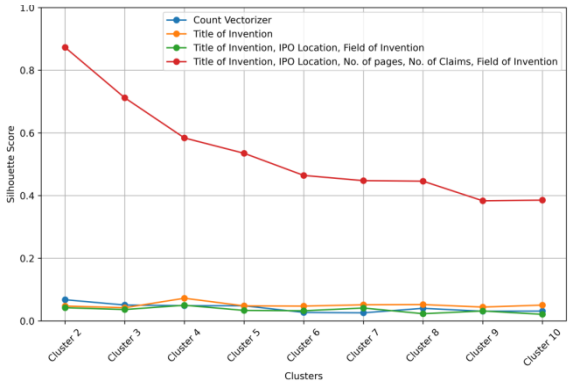

Fig. 8. Comparisons of different set of Features with K mean Algorithm for Count Vectoriser

Table 4. Silhouette score CountVectoriser with K-Means Algorithm

Sr. No.	Algorithm	Features	Silhouette score	Optimum Cluster Number
1	K-Means with Count vectoriser	Title of the Invention	0.0676	2
2		Title of the Invention, IPO Location	0.0724	4
3		Title of the Invention, IPO Location, Field of Invention	0.0500	4
4		Title of the Invention, IPO Location, No. of pages, No. of Claims, Field of Invention	0.8728	2

First Experiment (Titles of Inventions, 2 Clusters, Silhouette Score: 0.0676):
In the initial experiment, the clustering was performed exclusively based on the titles of inventions, resulting in two clusters. However, the silhouette score of 0.0676 indicates a weak separation between these clusters. Patent titles, being succinct and often standardized, might lack the complexity needed for robust categorization.

- Second Experiment (Titles of Inventions, IPO Location, 4 Clusters, Silhouette Score: 0.0724):
  The second scenario incorporated both the titles of inventions and IPO (Initial Public Offering) locations as features, expanding the clusters to four. Despite this augmentation, the silhouette score only marginally improved to 0.0724. While geographical information added diversity, it was still not sufficient for substantial enhancement in clustering quality.
- Third Experiment (Titles, IPO Location, Field of Invention, 4 Clusters, Silhouette Score: 0.0500):
  The third experiment included an additional feature: the 'Field of Invention.' Despite this addition, the silhouette score decreased to 0.0500, indicating that this categorical information did not significantly contribute to improved cluster separation. It suggests that the chosen categories might not be distinct enough for effective clustering in this context.
- Fourth Experiment (Titles, IPO Location, No. of Pages, No. of Claims, Field of Invention, 2 Clusters, Silhouette Score: 0.8728):
  The final and most comprehensive experiment incorporated multiple features: titles, IPO locations, structural details (number of pages and claims), and the categorical 'Field of Invention.' This approach yielded two highly distinct clusters with a remarkable silhouette score of 0.8728, indicating strong separation. The inclusion of various dimensions of data led to significant improvement, suggesting that a diverse and inclusive feature set is critical for effective clustering of complex datasets such as patents.
  The data clearly demonstrates the importance of feature diversity in clustering. Starting with basic textual information, the addition of geographic and structural details, along with categorical data, significantly enhanced the quality of clustering. It underscores the necessity of adopting a holistic approach that considers various facets of data for meaningful and insightful categorization of Indian patents. The final experiment, incorporating titles, geographic, structural, and categorical features, resulted in the most distinct and meaningful clusters, emphasizing the significance of a comprehensive feature set in patent data analysis and categorization.

6.2 Clustering with TF-IDF Vectoriser and K-Means

The results of clustering studies using K-means with TF-IDF Vectorizer, taking into account various feature sets, are shown in the accompanying table. Depending on which attributes are included, a measure of the quality of clustering is provided by the silhouette score. The high silhouette score of 0.8742 obtained with the set of features ('Title of the Invention, IPO Location, No. of pages, No. of Claims, Field of Invention'), leading to two optimal clusters, prominently illustrates how the inclusion of specific attributes like 'No. of pages' and 'No. of Claims' significantly improves clustering effectiveness. This result emphasizes the importance of choosing pertinent features and highlights how they affect

clustering success. According to the results, comprehending how invention titles, IPO locations, and other details like page numbers and claims interact might help reveal important trends that lie beneath the surface of the patent information. Through a deeper knowledge of the links between various patent features, researchers and practitioners can use these insights to support innovation analysis and strategic decision-making processes.

Table 5. Silhouette score results with TF-IDF Vectoriser and diffefent set of Features

Cluster Number	Title of the invention	Title of the Invention, IPO Location	Title of the Invention, IPO Location, Field of Invention	Title of the Invention, IPO Location, No. of pages, No. of Claims, Field of Invention
Cluster 2	0.0043	0.1891	0.1281	0.8743
Cluster 3	0.0060	0.2552	0.1698	0.7161
Cluster 4	0.0076	0.2948	0.1951	0.5961
Cluster 5	0.0088	0.1765	0.1424	0.5442
Cluster 6	0.0100	0.0841	0.1140	0.4633
Cluster 7	0.0101	0.0851	0.1340	0.4603
Cluster 8	0.0117	0.0853	0.1416	0.4578
Cluster 9	0.0097	0.0858	0.1477	0.4009
Cluster 10	0.0096	0.0859	0.1445	0.3997

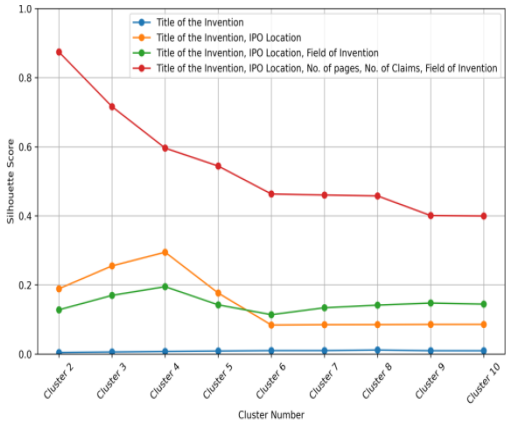

Fig. 9. Comparisons of different set of Features with K mean Algorithm for TF-IDF Vectoriser

Table 6. Silhouette score TF-IDF Vectoriser with K-Meanss Algorithm

Sr. No.	Algorithm	Features	Silhouette score	Optimum Cluster Number
1	K-Meanss with TF-IDF Vectoriser	Title of the Invention	0.0117	8
2		Title of the Invention, IPO Location	0.2947	4
3		Title of the Invention, IPO Location, Field of Invention	0.1956	4
4		Title of the Invention, IPO Location, No. of pages, No. of Claims, Field of Invention	0.8743	2

1. K-Meanss with TF-IDF Vectorizer (Title of the Invention) - 8 Clusters (Silhouette Score: 0.0117):

In this experiment, the TF-IDF vectorized representation of invention titles was used for clustering, employing K-Meanss with eight clusters. The resulting silhouette score of 0.0117 indicates weak separation and little distinction between the clusters. This suggests that clustering based solely on the textual content of invention titles, even with TF-IDF transformation, did not yield meaningful or distinct clusters.

K-Means with TF-IDF Vectorizer (Title, IPO Location) - 4 Clusters (Silhouette Score: 0.2947):

This experiment incorporated invention titles and IPO locations, both represented using TF-IDF vectors. K-Meanss clustering with four clusters was applied, resulting in an improved silhouette score of 0.2947. The inclusion of geographical context (IPO locations) led to better cluster separation, indicating that patents from different geographical regions tended to form more distinct clusters.

2. K-Means with TF-IDF Vectorizer (Title, IPO Location, Field of Invention) - 4 Clusters (Silhouette Score: 0.1956):

In this setup, the field of invention was included, represented as TF-IDF vectors. K-Means clustering with four clusters was applied, resulting in a silhouette score of 0.1956. The inclusion of the thematic content of patents (field of invention) provided a moderate enhancement in cluster quality. This suggests that considering the specific domain or industry of the patents improved the clustering effectiveness.

3. K-Means with TF-IDF Vectorizer (Title, IPO Location, No. of Pages, No. of Claims, Field of Invention) - 2 Clusters (Silhouette Score: 0.8742):

The final experiment integrated multiple features: titles, IPO locations, the number of pages, the number of claims, and the field of invention, all represented as TF-IDF vectors. K-Means clustering with two clusters achieved a significantly higher silhouette score of 0.8742. This substantial improvement indicates that this comprehensive feature set led to well-defined and distinct clusters, highlighting the importance of considering multiple aspects of patents for effective clustering.

The experiments underscore the critical role of comprehensive features, including textual content, geographical context, and structural attributes (such as page and claim numbers), in enhancing clustering quality. Clustering based solely on the textual content of invention titles (even with TF-IDF transformation) resulted in poor separation between clusters. The inclusion of additional contextual information, especially geographical and structural data, substantially improved the cluster quality. The optimal cluster number varied across experiments, emphasizing the importance of choosing an appropriate number of clusters for meaningful data segmentation. The last experiment, which incorporated multiple features, including both textual and structural attributes, resulted in highly distinct clusters, highlighting the significance of a diverse feature set for robust clustering.

6.3 Clustering with word2vec

Clustering with word2vec and K-Means

The silhouette scores presented for varying numbers of clusters (n clusters) provide valuable insights into the quality of clustering for the given dataset. The silhouette score measures how well-defined the clusters are within the data, with higher values indicating more distinct and cohesive clusters. Features are taken as Title of the Invention, IPO Location, No. of pages, No. of Claims, Field of Invention because these features show the highest silhouette score in previous analysis. Principle component analysis is done and obtained 2 principle components for analysis. In this case, the highest silhouette score of approximately 0.875 is achieved when the data is clustered into two groups. This indicates a strong separation between the clusters, with data points within each cluster being highly similar to one another and significantly different from points in the other cluster. A score above 0.7 generally suggests well-defined clusters. When the number of clusters is increased to 3, the silhouette score remains relatively high at around 0.716. While slightly lower than the two-cluster scenario, this score still suggests good cluster separation. Three distinct groups can be identified, although the cohesion within each cluster is not as strong as in the two-cluster case. As the number of clusters continues to increase to four, five, six, and beyond, the silhouette scores gradually decrease. This decrease indicates diminishing cluster quality. The clusters become less distinct and more overlapping, making it challenging to clearly separate the data points into meaningful groups. Specifically, when there are four clusters, the silhouette score drops to approximately 0.595, indicating a decrease in cluster cohesion and an increase in cluster overlap. Further increasing the number of clusters results in even lower silhouette scores, indicating weaker clustering structures. In summary, the analysis suggests that the data can be effectively grouped into either two or three clusters, with the two-cluster solution showing the strongest cluster separation. Beyond three clusters, the quality of clustering diminishes, leading to less interpretable and less meaningful groupings of the data. Therefore, selecting either two or three clusters would yield the most robust and insightful results for this dataset.

Table 7. Silhouette score results with word2vec Vectoriser and diffefent set of Features

Cluster Number	Title of the invention	Title of the Invention, IPO Location	Title of the Invention, IPO Location, Field of Invention	Title of the Invention, IPO Location, No. of pages, No. of Claims, Field of Invention
Cluster 2	0.0435	0.1726	0.1322	0.8751
Cluster 3	0.0402	0.2269	0.1722	0.7164
Cluster 4	0.0410	0.2630	0.1979	0.5954
Cluster 5	0.0162	0.1759	0.1512	0.5500
Cluster 6	0.0196	0.1001	0.1331	0.4644
Cluster 7	0.0227	0.0708	0.1429	0.4644
Cluster 8	0.0233	0.0670	0.1375	0.4633
Cluster 9	0.0230	0.0675	0.1423	0.4104
Cluster 10	0.0216	0.0537	0.1514	0.4156

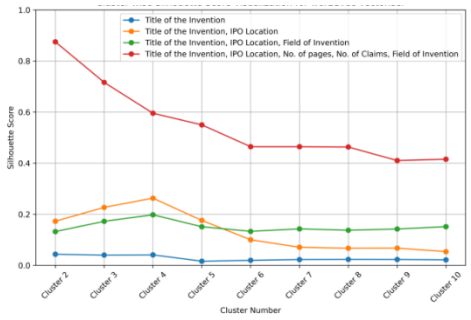

Fig. 10. Comparisons of different set of Features with K mean Algorithm for Word2vec Vectoriser

Table 8. Silhouette score Word2vec Vectoriser with K-Meanss Algorithm

Sr. No.	Algorithm	Features	Max. Silhouette score	Optimum Cluster Number
1	K-Means with Word2vec	Title of the Invention	0.0435	2
2		Title of the Invention, IPO Location	0.2630	4
3		Title of the Invention, IPO Location, Field of Invention	0.1978	4
4		Title of the Invention, IPO Location, No. of pages, No. of Claims, Field of Invention	0.8751	2

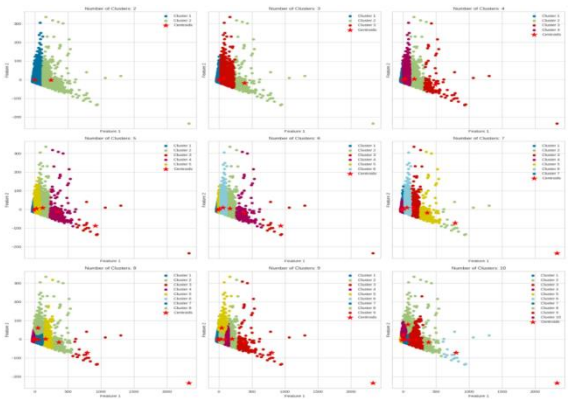

Fig. 11. Clusters wise scatter plot of cluster number 2 o 10 for word2vec

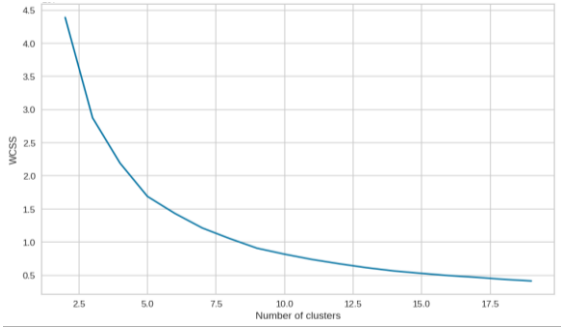

Fig. 12. Elbow method for finding optimum clusters Optimum numbers of clusters obtained are approximately 3 by elbow method and 2 by silhouette score method

Clustering with word2vec and Agglomerative Clustering (Hierarchical Clustering)

The features that display the greatest silhouette score in the preceding study are the Title of the Invention, IPO Location, Number of Pages, Number of Claims, and Field of Invention. After doing a principle component analysis, 2 primary components were discovered. The number of clusters that might be formed from 2 to 10 using the Silhouette Scores for clustering the provided data was calculated. The quality of the data points' clustering within each cluster is shown by these scores. From previous analysis word2vec method gives more accurate results hence for the further investigation of algorithms, the same word2vec method is used.

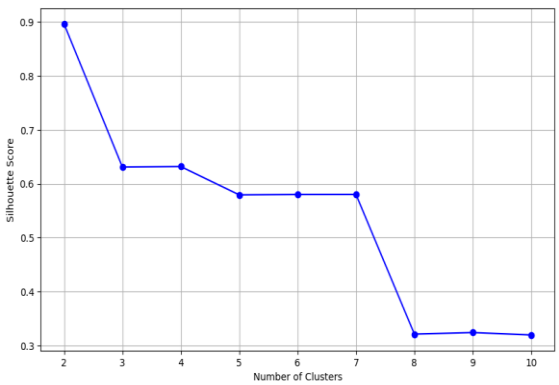

Fig. 13. Silhouette Scores for Different Cluster Numbers for Agglomerative Clustering with word2vec method

In this case of Agglomerative Clustering, the Silhouette Scores for 2 clusters (0.8965) significantly outperform those for higher cluster numbers, indicating a clear demarcation between the two clusters. As the number of clusters increases beyond 2, the Silhouette Scores gradually decrease, indicating that the clusters become less distinct and cohesive. Therefore, the optimal number of clusters for this dataset is 2. This implies that the data points exhibit strong similarities within their respective clusters, justifying the choice of dividing them into two distinct groups for the most meaningful analysis.

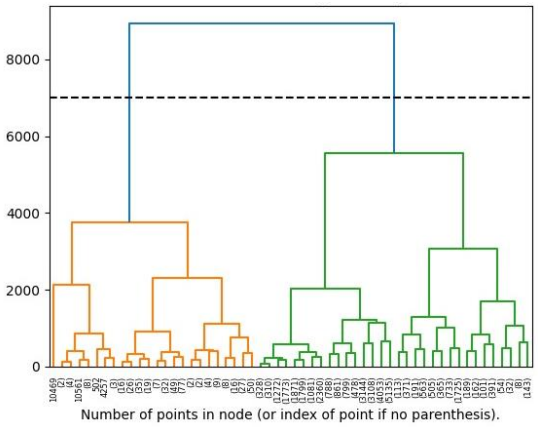

Fig. 14. Dendogram for Agglomerative Clustering

One of the primary applications of dendrograms is in determining the optimal number of clusters, a critical decision in clustering analysis. By visually inspecting the dendrogram and selecting an appropriate height to cut it, analysts can define distinct clusters. This method not only aids in finding natural groupings within the data but also helps in understanding the relationships between these groups. Additionally, dendrograms facilitate the comparison of different clustering algorithms, providing insights into their performance and suitability for specific datasets. When employing a dendogram, 2 clusters is the ideal number as well.

Table 9. Result Table for Best Algorithm and Technique

Sr. No.	Algorithm	Technique	Features	Silhouette Score	Optimum Clusters
1	K-Means	Count Vectoriser	Title of the Invention, IPO Location, No. of pages, No. of Claims, Field of Invention	0.8728	2
2	K-Means	TF-IDF Vectoriser	Title of the Invention, IPO Location, No. of pages, No. of Claims, Field of Invention	0.8742	2
3	K-Means	Word2vec	Title of the Invention, IPO Location, No. of pages, No. of Claims, Field of Invention	0.8750	2
4	Agglomerative Clustering	Word2vec	Title of the Invention, IPO Location, No. of pages, No. of Claims, Field of Invention	0.8965	2

In summary, Agglomerative Clustering with Word2Vec embedding produced the most distinct and well-separated clusters among all the techniques, highlighting its effectiveness in capturing intricate patterns within the data.

6.4 Percentage-wise Distribution of Patents in Different Fields for Each Cluster

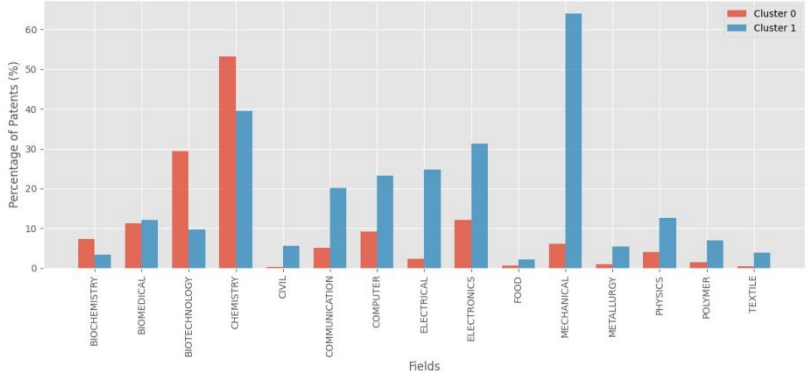

Fig. 15. Percentage-wise Distribution of Patents in Different Fields for Each Cluster

1. **Dominant Fields in Cluster 0:**

CHEMISTRY (36.99%): Cluster 0 has a significant concentration of patents in the field of Chemistry, indicating a strong focus on chemical innovations.

BIOTECHNOLOGY (20.47%): Biotechnology is another prominent field in Cluster 0, suggesting a substantial presence of biotechnological inventions.

ELECTRONICS (8.48%): The electronics field also shows a notable presence, although relatively lower compared to Chemistry and Biotechnology.

2. **Dominant Fields in Cluster 1:**

MECHANICAL (27.86%): Cluster 1 stands out with a high percentage of patents in the mechanical field, indicating a significant focus on mechanical engineering innovations.

CHEMISTRY (13.44%): Chemistry is also substantial in Cluster 1, although less dominant than in Cluster 0.

ELECTRONICS (10.77%): Similar to Cluster 0, Cluster 1 also has a noteworthy presence in the electronics field.

3. **Fields with Relatively Low Presence:**

FOOD, POLYMER, TEXTILE (Less than 1% each in both clusters): These fields show a minimal presence in both clusters, suggesting limited patent activity related to food technology, polymers, and textiles.

4. **Cluster Differences:**

Cluster 0 vs. Cluster 1: Cluster 0 appears to have a more diversified portfolio, with strong representations in Chemistry and Biotechnology. In contrast, Cluster 1 is significantly dominated by Mechanical innovations, although it also has substantial presence in Chemistry and Electronics.

5. **Fields with Potential Collaboration:**

CHEMISTRY: Given its significant presence in both clusters, Chemistry-related innovations could be a potential area for collaboration and knowledge exchange between Cluster 0 and Cluster 1.

6.5 Cluster-wise Distribution of NO_OF_PAGE, NO_OF_CLAIMS

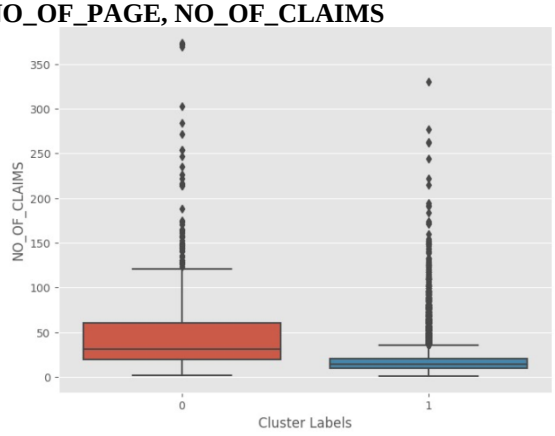

Fig. 16. Cluster-wise Distribution of NO_OF_CLAIMS

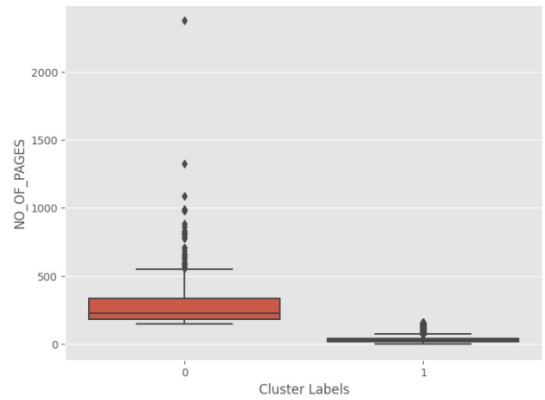

Fig. 17. Cluster-wise Distribution of NO_OF_PAGES

Cluster 0:
Average NO_OF_PAGES: 280.79
Average NO_OF_CLAIMS: 48.29

Cluster 0 represents a group of patents with a significantly higher average number of pages (approximately 281) and claims (around 48). Patents in this cluster tend to be extensive, containing comprehensive documentation and a substan-

tial number of claims. These patents likely cover complex innovations with detailed descriptions and multiple specific claims, indicating a higher level of complexity and depth in the inventions.

Cluster 1:

Average NO_OF_PAGES: 33.12

Average NO_OF_CLAIMS: 16.61

Cluster 1, in contrast, consists of patents with much shorter average lengths. The patents in this cluster have around 33 pages on average, indicating more concise documentation compared to Cluster 0. Additionally, these patents have approximately 17 claims on average, suggesting a focused scope of innovation. Patents in this cluster are likely to be more straightforward, with fewer pages and claims, indicating a more specific and targeted focus in their inventions.

Cluster-wise Complexity: Cluster 0 comprises patents with extensive content, indicating intricate and multifaceted innovations. Researchers and professionals analyzing patents in this cluster should be prepared for in-depth and detailed technical documentation.

Cluster-wise Conciseness: Cluster 1, on the other hand, represents patents with more concise content, suggesting simpler and more focused inventions. Professionals looking for patents with streamlined documentation and a narrower focus may find Cluster 1 more accessible for their research needs.

Understanding the average page and claim counts within each cluster aids researchers and analysts in choosing patents based on their complexity and depth, tailoring their investigations to the level of detail required for their specific research purposes.

6.6 Cluster wise Field of Invention

Cluster 0:

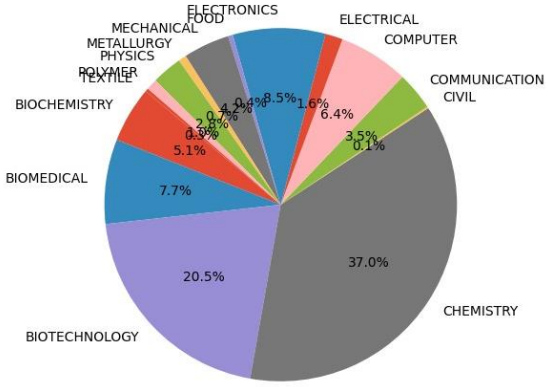

Fig. 18. Distribution of field wise Invention in cluster 0

Dominant Fields: In Cluster 0, the dominant fields are Chemistry (253 occurrences), Biotechnology (140 occurrences), and Electronics (58 occurrences). These fields have a substantial number of patents compared to other fields in this cluster.

Diversity: Cluster 0 exhibits a diverse range of fields, including Mechanical (29 occurrences), Biomedical (53 occurrences), and Communication (24 occurrences), showcasing a mix of technical and scientific patents.

Cluster 1:

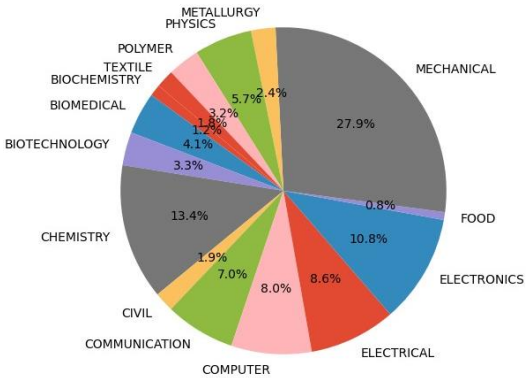

Fig. 19. Distribution of field wise Invention in cluster 1

Dominant Fields: Cluster 1 is notably dominated by Mechanical (9600 occurrences), making it the most prominent field in this cluster. Additionally, Chemistry (4632 occurrences) and Electronics (3711 occurrences) are significant fields in Cluster 1.
Specialization: Unlike Cluster 0, Cluster 1 is more specialized, with a strong focus on Mechanical patents, indicating a specific area of expertise or innovation in mechanical technologies.
In summary, Cluster 0 demonstrates a broader diversity of fields, while Cluster 1 is more specialized, particularly in Mechanical, Chemistry, and Electronics fields. The high number of patents in Mechanical in Cluster 1 suggests a specific area of innovation or research focus within this cluster.

7 Discussions

All analyses consistently found that the best number of groups for inventions was 2. This suggests two main categories among the inventions. Techniques understanding word meanings worked better for grouping inventions. Hierarchical clustering with Word2Vec was the most effective. One group had diverse fields like Chemistry, Biotechnology, and Electronics, while the other focused more on Mechanical Engineering. The first group had more complex patents, while the second had more focused innovations. These findings help organizations plan collaborations and understand patent trends for better research strategies.

Conclusion

The analysis showed that the inventions fall into two main groups, regardless of the method used. One group covers a wide range of scientific fields like Chemistry, Biotechnology, and Electronics, suggesting collaboration. The other group mainly focuses on Mechanical Engineering, indicating a specialized direction for deeper research. The patents in the first group have more pages and claims, indicating detailed and thorough inventions, while the second group's patents are more focused and concise. Cities like Delhi, Chennai, Mumbai, and Kolkata show different levels of innovation, with Delhi leading in patent filings. This highlights these cities as important hubs for inventions due to policies and support. Understanding these trends helps researchers focus their studies, guides policy-makers in planning, and assists organizations in collaborations and research planning for the future of inventions in India.

References

1. Cahyo, P. W., & Sudarmana, L. (2022). A Comparison of K-Means and Agglomerative Clustering for Users Segmentation based on Question Answerer Reputation in Brainly Platform. *Elinvo (Electronics, Informatics, and Vocational Education)*, *6*(2). https://doi.org/10.21831/elinvo.v6i2.44486
2. Da Poian, V., Theiling, B., Clough, L., McKinney, B., Major, J., Chen, J., & Hörst, S. (2023). Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry. *Frontiers in Astronomy and Space Sciences*, *10*. https://doi.org/10.3389/fspas.2023.1134141
3. Feng, S. (2023). Strategic Citations in Patents: Analysis Using Machine Learning. *Qeios*. https://doi.org/10.32388/vuk7qo.2
4. Gagliardi, I., & Artese, M. T. (2020). Semantic unsupervised automatic keyphrases extraction by integrating word embedding with clustering methods. *Multimodal Technologies and Interaction*, *4*(2). https://doi.org/10.3390/mti4020030
5. Grander, G., Ferreira da Silva, L., & Gonzalez, E. D. R. S. (2021). A Patent Analysis on Big Data Projects. *International Journal of Business Analytics*, *9*(1). https://doi.org/10.4018/ijban.288516
6. Imad Dabbura. (2018, September). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
7. Kalra, V., Kashyap, I., & Kaur, H. (2022). Improving document classification using domain-specific vocabulary: hybridization of deep learning approach with TFIDF. *International Journal of Information Technology (Singapore)*, *14*(5). https://doi.org/10.1007/s41870-022-00889-x
8. Kumar, A., Rajpal, A., & Rathore, D. (2018). Genre Classification using Word Embeddings and Deep Learning. *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018*. https://doi.org/10.1109/ICACCI.2018.8554816
9. Li, M. J., Ng, M. K., Cheung, Y. M., & Huang, J. Z. (2008). Agglomerative fuzzy K-Means clustering algorithm with selection of number of clusters. *IEEE Transactions on Knowledge and Data Engineering*, *20*(11). https://doi.org/10.1109/TKDE.2008.88
10. Li, S., & Gong, B. (2021). Word embedding and text classification based on deep learning methods. *MATEC Web of Conferences*, *336*. https://doi.org/10.1051/matecconf/202133606022
11. Lone, H., & Warale, P. (2022). Cluster Analysis: Application of K-Means and Agglomerative Clustering for Customer Segmentation. *Journal of Positive School Psychology*, *2022*(5).
12. Mehta, V., Bawa, S., & Singh, J. (2021). WEClustering: word embeddings based text clustering technique for large datasets. *Complex and Intelligent Systems*, *7*(6). https://doi.org/10.1007/s40747-021-00512-9
13. Mohammadi, I., Mehrabi, S., Sutton, B., & Wu, H. (2021). Word Embedding and Clustering for Patient-Centered Redesign of Appointment Scheduling in Ambulatory Care Settings. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2021*.

14. Mu, W., Lim, K. H., Liu, J., Karunasekera, S., Falzon, L., & Harwood, A. (2022). A clustering-based topic model using word networks and word embeddings. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00585-4
15. Munshi, M., Patel, M., Alqahtani, F., Tolba, A., Gupta, R., Jadav, N. K., Tanwar, S., Neagu, B. C., & Dragomir, A. (2022). Artificial Intelligence and Exploratory-Data-Analysis-Based Initial Public Offering Gain Prediction for Public Investors. *Sustainability (Switzerland)*, *14*(20). https://doi.org/10.3390/su142013406
16. Onan, A., & Toçoğlu, M. A. (2021). Weighted word embeddings and clustering-based identification of question topics in MOOC discussion forum posts. *Computer Applications in Engineering Education*, *29*(4). https://doi.org/10.1002/cae.22252
17. Reddy, M. V., Vivekananda, M., & Satish, R. U. V. N. (2017). Divisive Hierarchical Clustering with K-means and Agglomerative Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering. *International Journal of Computer Science Trends and Technology (IJCST)*, *5*(Sep-Oct).
18. Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using python. *International Journal of Innovative Technology and Exploring Engineering*, *8*(12). https://doi.org/10.35940/ijitee.L3591.1081219
19. Shri Rajesh Kumar Sharma, M. / S. / D. M. of C. and I. , D. for P. of industry and I. T. (2023). *Weekly Patent Application Granted during 2023* . https://data.gov.in/resource/weekly-patent-application-granted-during-2023
20. Shubham Chouksey. (2020, April 21). *Demonstrating Calculation of TF-IDF From Sklearn*. https://medium.com/analytics-vidhya/demonstrating-calculation-of-tf-idf-from-sklearn-4f9526e7e78b
21. Tache, A. M., Gaman, M., & Ionescu, R. T. (2021). Clustering word embeddings with self-organizing maps. Application on LaRoSeDa - A large romanian sentiment data set. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*. https://doi.org/10.18653/v1/2021.eacl-main.81
22. Tezgider, M., Yildiz, B., & Aydin, G. (2019). Improving Word Representation by Tuning Word2Vec Parameters with Deep Learning Model [Derin öğrenme Modeli ile Word2Vec Parametrelerinin Ayarlanarak Kelime Temsilinin Geliştirilmesi]. *2018 International Conference on Artificial Intelligence and Data Processing, IDAP 2018*.
23. Volodymyr Zhukov. (2023). *A Guide to Understanding Word Embeddings in Natural Language Processing (NLP)*. https://ingestai.io/blog/word-embeddings-in-nlp
24. Wahba, Y., Madhavji, N. H., & Steinbacher, J. (2020). Evaluating the Effectiveness of Static Word Embeddings on the Classification of IT Support Tickets. *Proceedings of the 30th Annual International Conference on Computer Science and Software Engineering*, *November*.