National College of Ireland

# "A Lightweight Semantic Patent Search System for the Electric Vehicle Domain"

MSc Research Project
MSc Data Analytics (MSCDAD_A)

Priya Sethi Khurana
Student ID: x23292547

School of Computing
National College of Ireland

Supervisor: Aaloka Anant

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | |
|---|---|
| **Student Name:** | Priya Sethi Khurana |
| **Student ID:** | X23292547 |
| **Programme:** | MSc Data Analytics **Year:** 2024-2025 |
| **Module:** | Research Practicum 2 |
| **Supervisor:** | Aaloka Anant |
| **Submission Due Date:** | August 11, 2025 |
| **Project Title:** | " A Lightweight Semantic Patent Search System for the Electric Vehicle Domain " |
| **Word Count:** | 8949(Including Reference) **Page Count** 23 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** Priya

**Date:** 11/08/2025

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# "A Lightweight Semantic Patent Search System for the Electric Vehicle Domain"

Priya Sethi Khurana
x23292547

**Abstract**

In technology fields like Electric Vehicles (EVs), where innovation is fast-paced and ideas often intersect, finding truly relevant patents can be challenging. This study introduces a tailored patent similarity search system designed to address that challenge by focusing on the semantic content of patent claims rather than surface-level keywords. Using a compact transformer model (msmarco-MiniLM-L-6-v3), combined with a lightweight claim-weighting approach, the system captures both the meaning and importance of different claims. These are indexed in a fast, scalable vector search engine (FAISS) for efficient retrieval. The system is designed to be both practical and easy to scale, showing strong performance in identifying closely related patents while filtering out irrelevant results. A series of experiments validate its effectiveness across both real-world and synthetic patent queries, supporting its role as a useful tool for researchers and IP professionals in fast-evolving domains.

# 1. Introduction

The growth of sustainable mobility such as Electric vehicles led to a increase in patent filings across similar domains such as charging, battery systems, energy management, motor control and controller of vehicle. In general patents are the valuable assets for an organization that legally protects innovation and work as a parameter that guides the R&D department as discussed by ('Industrial Policies and Innovation: Evidence from the Global Automobile Industry', no date). However, as this EV industry continues to grow, volumes and complexity of patents that are being filed also increases and it is impacting patents professional such as Patents analyst, patent examiner and individual researcher to find similar prior arts, ensuring non-infringement and monitoring activities of competitor is time consuming than ever before. To cater for this there is need of lightweight,

Also, regarding available solutions conventional patent searching tools work by searching keywords with a combination of Boolean operators such as AND, OR and CPC classes (Hain *et al.*, 2022). Although these solutions are simple and give results in large numbers and mismatched to original input due to lack of their semantic understanding as established in research by (Ali *et al.*, 2024)capabilities as they understand Full text as equal and do not capture the importance of semantic meaning embedded in patent claim (Helmers *et al.*, 2019). Mostly they consider both independent and dependent equally by often ignoring the fact that independent claims are more important both technically and legally (Brean, 2018). However, there is significant progress in advancement of research for this type of system such as transformer-based solutions that use NLP for significant improvement in patent retrieval such as BERT based models, some domain specific models that are trained on citation data and multi-level system that perform semantically as well. Most of them are computationally heavy processors not ideal for everyday use and most of them overlook importance claims as mentioned in research by (Wang *et al.*, 2020) and . By catering all these points, we are proposing a system that is advanced in terms of interpretation of retrieved results.

For understanding the research gaps in this area, a detailed conversation was conducted with experts in domain patent attorney, a researcher in IP analytics, and a startup founder. Based on that few issues were mainly highlighted in currently available solutions, and they lack reliable ranking, invention confidentiality when using AI based public Search platform and

on top of that most of them return results in great numbers that are not much related. Also, all of them show great interest in trying out an AI based patent search system that offers clear indicators of similarity for their search instead of outputs that show nothing as such.

This research caters for all these points by development of a lightweight, interpretable semantic similarity system which focused on Patent claims in EV domain. Also, this system focusses on the most informative part of the entire documents which is claims by giving more weightage to independent claims than dependent (using rule-based weightage step) rather than just focusing on complete patent document and metadata. This is mainly done to generate semantic embeddings using transformer-based models. The system uses FAISS based search approaches for fast and scalable retrieval of results together with clear explicit similarity scores. These scores allow users to assess not just *whether* two documents are related, but *how strongly* they align semantically. To strongly support both system development and clear evaluation the project introduced a similarity matrix that shows an effective and easy for understanding the similarity concept by showing effective alternative of traditional evaluation such as precision or recall value which shows outputs of model not output of complete holistic system.

**Research Aim and Objectives**

Also, using the given technology this system answers following research question

*How can a lightweight and interpretable semantic similarity system be designed using claim-level weighting to improve patent retrieval accuracy and enable practical, score-based evaluation for professional users?*

Also, to address the research questions this research has following objectives

- **Investigate**: To evaluate the current state of semantic similarity approaches in patent retrieval, particularly at the claim level.
- **Design:** To design a scalable, transformer-based framework that emphasizes claim-level semantic embeddings and fast retrieval.
- **Implement:** To implement the system using FAISS for efficient vector search, prioritizing interpretability and real-time usability.
- **Evaluate:** To evaluate the system using a newly developed similarity score matrix to measure relevance, ranking stability, and confidence.

As a result, this study contributes to generation of a claim focused, efficient and explainable patent similarity search system particularly focusses on evolving domain like Electric vehicles innovation. It mainly targets real concerns highlighted by experts in domain, mainly transparency, usability and decision making while providing a generalizable framework for scalable semantic retrieval. This research is designed for trustworthy and efficient patent analytics in real world.

# 2.Related Work

In the domain of finding similar patents, recent research now doesn't focus on conventional keyword-based search systems rather than exploring advanced, semantically driven approaches. The detailed review divides similar and relevant studies among categories such as embedding strategies, claim-level modelling, hybrid evaluation methods, and interpretable similarity metrics that directly inform and position the current research.

**2.1 Embedding Models for Patent Similarity:** Few studies that compares transformer based embedding works as the base semantically patent retrieval by (Ascione and Sterzi, 2024) conducted a comprehensive comparison between traditional vector representations and

transformer-based models such as SBERT, demonstrating that domain-adapted SBERT variants consistently outperform baselines in identifying relevant prior art. Most important their study emphasized the variation between computational efficiency and semantic depth, justifying lightweight embedding choices like MiniLM in resource-conscious systems.

Similarly, (Bekamiri, Hain and Jurowetzki, 2021) introduced PatentSBERTa, a BERT-based sentence embedding model fine-tuned specifically for patent claims. Their works shows how semantically rich representations support both similarity tasks and CPC classification, while offering transparent outputs for interpretability.

Also,(Hain *et al.*, 2022) extended this approach by using nearest-neighbor search over large-scale embeddings to map patent-to-patent similarity networks, applying them to electric vehicle (EV) patent datasets and validating their model through citation co-occurrence and knowledge flow.

To sum up these taking all of the above as baseline my research targets to implement a lightweight yet semantically aware retrieval approach, optimizing performance for practical use without large-scale fine-tuning

### 2.2 Claim-Level and Structural Similarity Modelling

As already evident that independent claims mostly discloses the novelty of a patent, considering this recent work by (Vowinckel and Hähnke, 2023) targets claim-level granularity. This particular proposed SEARCHFORMER, a Siamese transformer model trained on citation triplets to better capture claim-to-claim semantic alignment. Their results showcase improvements over traditional keyword or full-text comparisons.

Together with this(Matthias, Heidari and Hewel, 2024) further improved this by dividing claims into chunks by assigning them weights and computing a similarity score. This chunk-level approach reflects an evolution toward deeper claim structure modelling.

A study by (Yoo, Xu and Cao, 2025) introduced Patent Mind which is a reasoning based solution that works on multiple aspects by dividing into parts such as claim scope and technical intent. This study works as a base for my work on conceptual level by highlighting the need of transparent, interpretable similarity aligned with how domain experts evaluate novelty scoring.

Taking these as motivation my system emphasizes claim-level weighting by specifically highlighting independent claims for generation more meaningful similarity scores aligned with patent structure.

### 2.3 Hybrid and Multi-Modal Evaluation Methods

Also, for this many researchers have shared multi-model approach to improve evaluation. Mainly a study by (Yoo *et al.*, 2023) proposed a hybrid model combining semantic text, IPC code co-occurrence, and bibliographic features to compute similarity through a weighted distance function.

And, a study by (Zhu *et al.*, 2023) introduced a multi-dimensional fusion strategy by incorporating text embeddings, dot-matrix comparison, and syntactic pattern weighting to rank patents by disclosure similarity.

Together with this , fine-tuned SciBERT models trained using triplet networks have been explored by (Acikalin and Kutlu, 2022) , further validating the utility of structure-aware models in capturing latent similarity.

In comparison to these, this study avoids multi-source fusion and focuses on a streamlined yet interpretable model, relying on embedded semantic claims and internal score structures.

### 2.4 Evaluation Using Similarity Scores

A key distinction of this research lies in its use of raw similarity scores to evaluate system effectiveness, avoiding the need for binary relevance labels. And, study by (Helmers *et al.*, 2019) Helmers et al. (2019) demonstrated how similarity scores between claim embeddings could be thresholded to rank prior art candidates and assessed using ROC/AUC metrics. This aligns closely with my introduction of a score-based evaluation matrix.

This research extends this basic idea of similarity scoring by a full similarity matrix with quantitative thresholds and user-oriented metrics to enhance both evaluation interpretability and trust on solution as it will show clear indication of the similarity of search query.

This literature review shows how the current research is taking motivation from the strong prior work while offering a distinctive contribution through its focus on claim-level weighting, CPU-efficient architecture, and score-based interpretability.

# 3. Methodology

This section Shows the details of the technical and systematic steps involved for the development and evaluation of a lightweight interpretable semantic patent similarity system specially for the Electric vehicle (EV) domain. The proposed methodology integrates knowledge captured in NLP and sematic search and addressing limitations of existing patent retrieval system which is ignoring importance of claims and lack of transparency in similarity scoring.

The methodology was inspired by structured analytical frameworks such as CRISP-DM and KDD particularly in their emphasis on systematic data preparation, model building, and evaluation the methodology is not strictly focusing on either of the approach. The mentioned steps and approach were chosen based on the domain needs in the EV sector, reproducibility principles, and efficiency considerations common in Natural language processing pipelines (e.g., text preprocessing, embedding, and evaluation). This resulted in efficient pipeline that included claim-level preprocessing, semantic embedding, FAISS-based retrieval, and score-based evaluation designed for interpretability and computational efficiency.

The Methodology section includes following steps: (1) preprocessing and claim-level preparation of EV patent data; (2) application of a custom claim weighting heuristic to emphasize legal and technical significance; (3) use of semantic embeddings to capture the meaning of each claim; and (4) implementation of a FAISS-based similarity search system paired with an interpretable evaluation framework based on similarity scores. These integrated steps form an efficient pipeline which is computationally efficient and is reproducible that helps in more accurate and explainable retrieval of semantically related prior art.

### 3.1 Data Collection and Preprocessing:

For this step data was sourced from the Google Biq Query via a USPTO public Dataset which offers access to large patent metadata and claims text as well. The dataset was filtered using keyword and Cooperative Patent Classification (CPC) codes to pick documents related to EV domain filled in US and published in last 10 years. As a result of this process total 12,000 patents were extracted for data corpus. Also, the Ev domain is selected for my research due to its high innovation activity and increasing complexity of patents as technologies are evolving and for this particular reason my study will work as highly valuable contribution in semantic patent searching domain.

Also, for this stage which is Data preprocessing following steps were followed to prepare data for modelling steps:

- **Missing Value Handling:** For this step we have checked Records with missing or blank entries in key fields (title, claims, publication_number) were removed. These fields are essential for semantic embedding and identification. And if we didn't apply this step incomplete records in dataset could cause downstream embedding errors or inconsistencies in indexing. This step ensures data reliability and structural uniformity for claim-level processing.
- **Text Cleaning:** For this step all textual fields were lowercased, and excessive whitespaces or special characters were removed using regex. The purpose of this step to ensure that no unnormalized text could hinder semantic embedding by introducing inconsistent patterns. For example, the same phrase with varying punctuation or casing might be embedded differently. Text cleaning helps reduce noise and ensures embedding reflects real semantic content.
- **Exploratory Analysis:** At this stage Basic statistics such as the distribution of claim counts were generated, and visual tools like histograms were used to understand data structure. This step was applied for better understanding the structure and variation in data for example how many claims. how many claims to expect, whether padding or truncation is needed, and whether claim count influences semantic relevance.
- **Claim Splitting:** For this A regular expression-based method split long claim strings into a structured list using numbering patterns (e.g., 1., 2.). Claims shorter than 20 characters were discarded. The main purpose of this step is mostly in all patent databases; all claims are stored together in one large block of text. This makes it difficult to analyses or compare individual claims, which is how patent experts typically assess novelty or scope. By splitting the text into separate claims, we can process each one independently, assign it a weight based on its importance, and compute more accurate similarity scores. This step is essential to move from a generic text-based model to a structured, claim-aware system that shows how patents are examined in real-world legal and technical contexts.

As a result, after applying multi-stage data preprocessing and cleaning which includes removing entries which have missing claims or has special character, we have obtained a clean dataset. And we have kept records which have fully defined title, claims, and publication number this ensures consistency for downstream semantic modelling. On this basis the final dataset, drawn from a larger electric vehicle-related patent corpus, was ready for weightage and semantic claim processing and embedding.

### 3.2 Claim Weighting

This step is considered as part of this pipeline as it is known that all Patent claims including the USPTO are not all equally informative or significant Independent claims often define the broadest legal protection(Brean, 2018), while dependent claims offer narrower details which are mostly supplementary . This legal and semantic distinction has been applied in recent literature, where claim-level modelling has emerged as a primary focus (Yoo, Xu and Cao, 2025). Also, supporting this view, legal-economic studies show that independent claims are more influential in defining novelty, valuation, and enforceability.

And for highlight and leveraging claim importance we have implemented claim weighting method where each claim is assigned numerical score based on the following factor:

- Its position in the list meaning the earlier claims is weighted more than the later one, by using following formula

$$\text{position\_score}_i = \max(1.0 - 0.05 \times i, 0.3)$$

  The score linearly decreases by 0.05 for each subsequent claim index * $i_1$, but never falls below a minimum threshold of 0.3. This ensures that dependent claims still contribute meaningfully, without dominating the semantic signal.
- Its textual length and longer claims are more Descriptive based on the length of the claims and that's mostly applied to independent claims and is computed based on the following formula

$$\text{length\_score}_i = \min(500\text{len}(\text{claim}), 1.0)$$

  In this Claims are scaled based on their character length, with a cap at 500 characters. Any claim longer than 500 characters is given a maximum length score of 1.0.

And, the final weight is average of both factors, which is slightly biased towards length for semantic value using following formula

$$w_i = 0.6 \times \text{length\_score}_i + 0.4 \times \text{position\_score}_i$$

According to this a weighted average is used, with 60% emphasis on length and 40% on position. This balance shows that detailed claims are slightly more informative than placement alone, particularly when claims are auto extracted without manual tagging. Recent empirical studies by (Ragot Sébastien) show that longer independent claims correlate positively with key indicators of patent value, such as forward citations and maintenance rates. This justifies design choice for this study to assign greater weight to longer claims, assuming that they have more significant aspects of the invention.

This step and design help system to generate embedding that give comparatively greater attention to the most meaningful part of the patent which satisfy both legal aspect and goals of semantic Patent retrieval system.

### 3.3 Claim Embedding using Transformer Models

After preprocessing and applying weightage to claims with the purpose of giving higher value to independent claims, the next phase includes transformation of each patent claims claim into a dense, fixed-size vector representation(embeddings) using transformer-based sentence embedding models. This transformation was essential for enabling semantic similarity comparison across patents which is important part of retrieval system

For this step, the sentence-transformers/msmarco-MiniLM-L6-v3 model was employed. This model is a compact variant of BERT developed by Microsoft and optimized for sentence-level semantic similarity tasks. It was trained on the MS MARCO dataset, which is widely used in information retrieval research. One of the primary motivations for selecting this model was its balance between performance and efficiency. Also, it achieves competitive results on semantic textual similarity benchmarks, and it also has a relatively small memory consumption and fast inference times, making it well-suited for CPU-based environments and lightweight deployment which is one of the goals of this project.

The choice of MiniLM over larger models such as BERT or RoBERTa was intentional. MiniLM models are significantly more resource-efficient while retaining competitive accuracy. According to study by(Wang *et al.*, 2020) MiniLM achieves up to 99% of BERT's performance on natural language understanding tasks but with fewer parameters and lower latency, which aligns with the project's goal of building a lightweight yet accurate system. The *msmarco-MiniLM-L-6-v3* as discussed in (*MSMARCO Models — Sentence Transformers documentation*, no date) variant also supports high-throughput inference approximately

18,000 queries/sec on CPU and 750 queries/sec on GPU making it ideal for scalable and cost-efficient applications.

In this step each claim in the dataset was passed through the MiniLM model to generate a 384-dimensional dense embedding. However, since a patent typically contains multiple claims, the model needed to compute a single vector representation per patent. To achieve this, a **weighted aggregation** of the claim embeddings was applied. The weights were derived from the scoring mechanism discussed in Section 3.2, which shows significance of both the position and length of each claim. Specifically, more importance was given to earlier and longer claims, based on legal literature by (Brean, 2018) and domain insights suggesting that independent and detailed claims carry greater significance in defining patent novelty and scope.

This method and approach is responsible for semantically rich representations of patents that depend on internal patent claim structure, and it is a distinct feature in comparison to prior art approaches used in similar literature that mostly rely on unweighted textual text. And these generated embeddings were used for similarity score computation and indexing that is essential for forthcoming patent retrieval step.

### 3.4 Similarity Indexing and Retrieval using FAISS

Patent retrieval system which has similarity as parameter requires both semantically meaningful representations and highly efficient mechanism to search across thousands of such vectors in real time. To address this, my research uses FAISS (Facebook AI Similarity Search) which is an open-source vector database library developed by Meta AI that enables scalable and efficient similarity search over high-dimensional embeddings.

For efficient semantic search over patent embeddings, we have used the FAISS library which is a widely used vector similarity toolkit developed by Meta AI. FAISS offers exact search methods like IndexFlatIP, which supports inner-product comparisons ideal for cosine similarity. According to (Douze *et al.*, 2025) FAISS's design balances scalability and precision while preserving full nearest-neighbor accuracy The original FAISS introduction as discussed by ('Faiss: A library for efficient similarity search', 2017) also demonstrates its suitability for billion-scale vector retrieval and highlights advantages in memory usage and speed over traditional search platforms.

After the embedding step discussed in section 3.3 each patent's claim set into a 384-dimensional vector using the MiniLM model, the resulting vectors are stored in a dense vector index. Also, FAISS supports fast nearest-neighbour search by leveraging optimized indexing structures, that include flat indexes, quantization, and GPU acceleration. In this work, the **IndexFlatIP** (Inner Product-based flat index) is used, which is particularly suited for cosine similarity when embeddings are normalized.

This approach provides three key advantages:

- **High-throughput search**: FAISS allows retrieval of the top-k similar items for any query vector in milliseconds, making it ideal for scaling up to large patent dataset
- **Accuracy without approximation**: It includes IndexFlatIP that performs an exact inner product search, avoiding approximation errors, which is crucial for legal contexts like patent similarity.
- **CPU-efficient and deployable**: This system uses normalized embeddings and lightweight models (MiniLM), FAISS operates efficiently even on standard CPUs following the project's goal of a lightweight yet accurate tool.

For ensuring data quality in this case, only vector with valid 384-dimensional embeddings are retained in the final FAISS index. In this each query claim set is embedded using the same logic (claim weighting + MiniLM) and compared against the corpus to obtain top-k matches. And, the similarity score is computed in the form of an inner product that reflects the true semantic closeness of the query patent to the indexed ones. This stage effectively serves as the retrieval component of the system, and forms the core of the ranking-based evaluation matrix implemented later.

### 3.5 Similarity Score Computation and Matrix Construction

Based on FAISS-based indexing of claim embeddings, the next step is computing semantic similarity scores between each input query and the pre-embedded dataset. Although, FAISS function as the engine for efficient vector retrieval, this stage uses it to extract ranked similarity scores for further evaluation of system in real case scenario. The main goal of this is to enable a structured, interpretable comparison of how closely each result aligns with the input patent query which is either text of any patent claims which is not from main corpus or any text of any invention in EV domain.

To understand this concept practically, the query embedding captured using the MiniLm model and weighted claim is compared to each vector in FAISS index using the cosine similarity (Normalized). And, for this the system returns the top K most similar patent results and a similarity score. This score is then framed into similarity matrix where each row corresponds to a query and each column to a matched result from the patent corpus.

This novel matrix serves as a method that allows transparent evaluation & ranking based analysis. It also includes computation of values such as top-1 similarity, mean similarity across top 10 results, and score drop-off patterns and this shows the details of the matches. This approach serves as the need for legal domains where graded relevance preferable rather than binary labels, that helps user to interpret not just what was retrieved but how similar they are importance of each result.

| Metric | What it Captures | Alignment with System Goals |
|---|---|---|
| Top-1 Similarity Score | Measures how semantically similar the **best match** is for each query. | Indicates whether the system retrieves at least one **highly relevant** document per query. |
| Mean Top-10 Similarity Score | Assesses the **overall relevance** of all retrieved patents, not just the top result. | Ensures the system provides **consistently relevant results**, supporting robust decision-making. |
| Drop from Top-1 to Mean | Quantifies the **ranking stability** by comparing the top hit to the rest of the list. | Helps determine if retrieval quality is spread evenly or dominated by a single strong match. |
| Count with Similarity ≥ 0.6 | Counts how many retrieved results are **moderately relevant or above.** | Reflects the system's ability to return a **broad set of relevant patents**, not just individual hits. |
| Count with Similarity ≥ 0.7 | Identifies the presence of **strongly relevant matches** per query. | Indicates the system's success in identifying **semantically meaningful links.** |

| | | |
|---|---|---|
| Count with Similarity ≥ 0.8 | Tracks the frequency of **very high similarity matches.** | Demonstrates how often the system retrieves **near-duplicate-level matches**, supporting high precision needs. |

**Table 1: Similarity Scoring Matrix**

This scoring framework showcases a **more interpretable and continuous view** of retrieval quality compared to traditional binary relevance assessments. This matrix structure also facilitates comparisons across future experiments and helps identify trends such as over-reliance on a single match or inconsistency across result ranks. It also prepares the foundation for later experiments, model comparisons, and domain-specific evaluations, as described in the subsequent chapters.

Also, while traditional binary relevance metrics as discussed by (Ali *et al.*, 2024) , which judge only whether a document is relevant or not and similarity score matrix in this study uses a ranking-aware perspective where both position and strength of semantic similarity scores are considered. This provides more efficient understanding of retrieval performance that captures not just if relevant results exist, but how well they are ranked and distributed in the results list.

# 4. Design Specification

This chapter represents a thorough specification of the overall architecture of system which includes functional components and design decisions which are foundation of patent similarity search tool developed in this study. The purpose of this system is to enable meaningful and semantically focused comparison between patents, and this holds extra value in domains like EV (Electric Vehicle) where technology is evolving faster and so the chances of overlapping claims.

The design of the patent similarity was developed when there was need of scalable, interpretable, and domain-relevant retrieval framework. This chapter shows the core design specifications including the characteristics of the dataset (Section 4.1), the modular architecture of the system (Section 4.2), and the specific algorithms and models selected (Section 4.3). In combination these components reflect a focus on lightweight and semantically valuable processing of patents claims where special attention is given to legal structure, embedding techniques, and vector-based search. Also, each design decision was motivated by prior research, real-world constraints, and the unique nature of Electric Vehicle (EV) patent datasets that ensures that system is technically sound and practically applicable.

This chapter provides a step-by-step detail of the system's key modules, emphasizing the 'what', 'how', and 'why' of each design choice to establish a clear idea of the system's intended behavior and capabilities. This is crucial not only for reproducibility but also to justify the model's design in the context of broader research and industrial applicability.

## 4.1 Data Specification

The dataset used in this project was sourced from the USPTO Data Repository, accessed via Google Big Query, which offers scalable querying of public patent records. A filtered subset of ~12,000 utility patents was curated based on relevance to Electric Vehicle (EV) technologies, using keyword-based search terms (e.g., "EV battery", "charging", "electric motor", etc.) and CPC codes across the title, abstract, and claims fields.

Each patent entry in the dataset included the following key fields:

Publication number, title, abstract, claims and together with this following fields were added after Model application claim list, claim weights, sbert_embedding

Also, for making data suitable for semantic similarity system several preprocessing steps were performed:

- **Missing Values Removed**: It is to ensure that patents missing core fields (title, claims, publication number) were dropped.
- **Text Normalization**: In this all text was lowercased, unwanted characters removed via regex, and whitespace normalized.
- **Claim Splitting**: For these steps claims were split into structured lists using regular expressions to isolate numbered claims (e.g., "1. …", "2. …").
- **Claim Weighting**: At this step each claim received a calculated weight based on its position and length.

The primary reason for choosing Electric vehicle related patents as a field aligns with the objective of generating a domain-relevant tool where semantic similarity is particularly important due to technical overlaps and as shared in report in article (Insights, 2025) where frequent evolving innovation in the EV field is driving more patent filings. The cleaned and embedded dataset offered a scalable and interpretable foundation for downstream tasks such as similarity matching, relevance ranking, and UI-driven retrieval.

### 4.2 System Architecture

The core architecture of the proposed patent similarity system is designed to be modular, scalable, and explainable. It has multiple components which are data preprocessing and semantic embedding to vector indexing and evaluation and each of them carefully selected to address the challenges of efficient prior-art retrieval in the Electric Vehicle (EV) domain.

- **Data Layer**: This layer includes the cleaned and pre-processed corpus of ~12,000 US EV-related patents. Preprocessing operations (detailed in Methodology) ensure consistency and enable downstream tasks like semantic parsing and claim-level processing.
- **Semantic Processing Engine**: This layer integrates claim splitting, weighting, and embedding. At this step claims are split and prioritized using a weighting mechanism and then passed through the MiniLM-based encoder. The design intentionally separates legal-context logic (claim weighting) from semantic embedding to allow independent optimization or replacement of either module.
- **Similarity Index Layer**: This layer is implemented using FAISS, this component stores the vector representations of patents and allows for fast nearest neighbour search. Its choice is motivated by the need for CPU efficiency and inner-product–based similarity.
- **Retrieval & Ranking Interface**: This module performs similarity search against the FAISS index for a given query, and ranks the results using a ranking-aware scoring matrix. The scores reflect not only presence of relevance but relative closeness of patents in the semantic vector space.
- **User Interface Layer**: The frontend was implemented using Stream lit, where end-users can input custom queries or explore test cases. The. parquet format allows efficient local loading and enables efficient retrieval of results.

For implementation of this system each of the layer is different from each other, that enables flexibility in using different models or may be advanced deployment environment for real time dashboard generation. This architectural setup ensures that system can be scalable to different domain and even more big patent dataset for retrieval of patents using intelligent frameworks.

### 4.3 Algorithm Specification

This section explains the core algorithms and reason behind model choices that works as backbone of the semantic similarity framework developed in this study. The system design integrates natural language processing and vector-based retrieval techniques to support a legally informed comparison of patent claims. Each algorithm or model was selected based on its computational efficiency, semantic robustness, and suitability for domain-specific patent analysis

The core algorithmic pipeline of the proposed system integrates a combination of semantic text embedding and dense vector search to retrieve patents with high conceptual relevance. At the base of the system lies the msmarco-MiniLM-L-6-v3 model, which is Sentence Transformers type, it is chosen for its balance of computational efficiency and semantic accuracy. Patent claims are preprocessed, split, and weighted before being individually embedded using this transformer model. The embeddings of these claims are then aggregated using a weighted average to form a single vector representation per patent. For similarity-based retrieval, FAISS (Facebook AI Similarity Search) is used to construct a vector index using inner product search, enabling efficient and scalable nearest neighbor retrieval across thousands of embeddings. This combination of light-weight embedding, weighted semantic representation, and fast retrieval architecture supports the system's goal of delivering interpretable and high-quality results for the Electric Vehicle patent domain.

# 5. Implementation

This chapter tells about the final implementation of the semantic patent similarity system developed in this study. Whereas the design specifications in Chapter 4 shows the conceptual framework, this section focuses on how the system was practically built, integrated, and tested using real patent data. The implementation process included dataset preparation, embedding generation, indexing, similarity scoring, and a front-end user interface to interact with the system. Each component was developed with careful attention to modularity, scalability, and interpretability ensuring that the system remains lightweight, efficient, and aligned with real-world applicability. The implementation was carried out using Python in a Google Collab environment and later transitioned to a local interface via Stream lit for demonstration and usability testing.

### 5.1 Overview of System Implementation

During development of the system, we followed a modular and pipeline-based approach. The process began with the extraction and preprocessing of 12,000 Electric Vehicle (EV)-related patents sourced from the USPTO dataset, ensuring a domain-specific focus. Each patent's claim text was cleaned and then split into individual claims based on numbering patterns, allowing for more granular semantic representation.

Following this, a custom weighting scheme was applied to each split claim to account for its structural significance and length, helping to prioritize more informative content during embedding. The claims were then semantically embedded using the msmarco-MiniLM-L-6-v3 transformer model, selected for its lightweight architecture and ability to capture deep contextual relationships efficiently. These embeddings were further normalized and indexed

using FAISS (Facebook AI Similarity Search), enabling high-speed retrieval of similar documents based on cosine similarity in vector space.

The user interface was developed using Stream lit to enable interactive querying and visualization of results. This included entering query claims, triggering a semantic search over the indexed corpus, and retrieving top-ranked patents with similarity scores. The entire pipeline from text preprocessing to FAISS-based retrieval was integrated seamlessly to ensure that user inputs could be processed in real-time without compromising system efficiency.

Throughout the implementation phase, the focus remained on building an end-to-end system that is not only technically sound but also usable by patent analysts, researchers, or domain experts. The underlying architecture supports scalability, while the transparency of each module allows for future upgrades, such as alternative embedding models or ranking strategies, with minimal changes.

### 5.2 Dataset and Ethical Considerations

This study uses a dataset of approximately 12,000 publicly available patent documents from the USPTO, accessed via Google Big Query. The dataset focuses on electric vehicle (EV) technologies and includes titles, abstracts, and detailed claim texts. A thorough preprocessing pipeline was applied to clean missing values, normalize text, and split claims into structured lists for further analysis.

Ethically, the project uses only open data, with no personal or sensitive information involved. All processing steps are transparent and reproducible, ensuring alignment with responsible AI research practices. The system was developed for academic purposes, with no commercial intent or proprietary data usage.

### 5.3 Tools and Technology Stack

The system was implemented using Python due to its wide ecosystem of data processing and machine learning libraries. Development was primarily conducted on Google Colab for rapid experimentation and later deployed via a Stream lit-based local UI for interactive exploration.

Key libraries and tools include:

- **Pandas & NumPy** for data handling and preprocessing
- **Regex** for custom claim text splitting
- **Sentence-Transformers** with the **msmarco-MiniLM-L-6-v3** model for semantic embeddings
- **FAISS** for efficient vector indexing and similarity search
- **Matplotlib & Seaborn** for exploratory data analysis
- **Parquet** format for storing processed data for fast I/O during UI usage

The overall stack was chosen for its **lightweight performance**, ease of experimentation, and suitability for scalable document similarity tasks.

# 6. Evaluation

This chapter presents a comprehensive evaluation of the semantic patent similarity system developed in this study. The primary objective of the evaluation is to assess how effectively the system retrieves patents that are semantically like a given query patent claim. Also, unlike conventional binary relevance assessments, which only judge whether a retrieved document

is relevant or not, this study adopts a ranking-aware evaluation approach that reflects graded relevance using similarity scores. The evaluation strategy is constructed around a set of experiments designed to test different aspects of the system's performance, including real-world applicability, robustness, and the effect of internal design choices. Each experiment aligns with the project's broader goals of enabling lightweight, domain-specific, and semantically meaningful patent retrieval in the Electric Vehicle (EV) domain.

### 6.1 Experiment 1: Real Query Evaluation

To assess how accurately and consistently the system retrieves semantically similar patents when provided with real EV patent claims not seen during training or embedding.

**Methodology:**
110 query patents were selected from the EV domain; each processed through the system to retrieve the top-10 most semantically similar patents from the FAISS-indexed corpus. The queries were embedded using the same weighted MiniLM method as the training corpus, ensuring methodological consistency. "In this context, a *query* refers to the structured, cleaned claim text of an EV-related patent used to retrieve similar entries from the main corpus." As a result of these following metrics scores are computed

| Metric | |
|---|---|
| **Number of Queries** | 110.000000 |
| **Top-K Retrieved** | 10.000000 |
| **Avg. Top-1 Similarity** | 0.733484 |
| **Avg. Mean Top-10 Similarity** | 0.659918 |
| **Avg. Drop (Top1 - Mean10)** | 0.073566 |
| **Avg. Count >= 0.6** | 7.318182 |
| **Avg. Count >= 0.7** | 3.363636 |
| **Avg. Count >= 0.8** | 0.681818 |

**Figure 1: Similarity score Matrix computed**

For each query, the system returned the Top-10 most semantically similar patent claims, and various ranking-aware metrics were computed to assess the quality and consistency of these results

- On the basis of below Figure 2 and metric table The ***Top-1 similarity*** score across all queries averaged **0.733**, indicating that the most similar result for each query was highly relevant. This is a strong indicator that the system's embedding and claim-weighting strategy accurately preserves legal and semantic meaning.

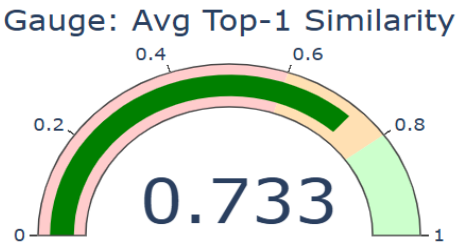


**Figure 2: Average Top-1 Similarity**

- A per below Figure 3, ***The Mean@10 similarity*** — average of all Top-10 retrieved claims — was 0.660, confirming that relevance was not limited to just the top result but extended across the full ranked list. This suggests robustness in retrieval beyond the first hit. (Use: Histogram showing Mean@10 distribution)

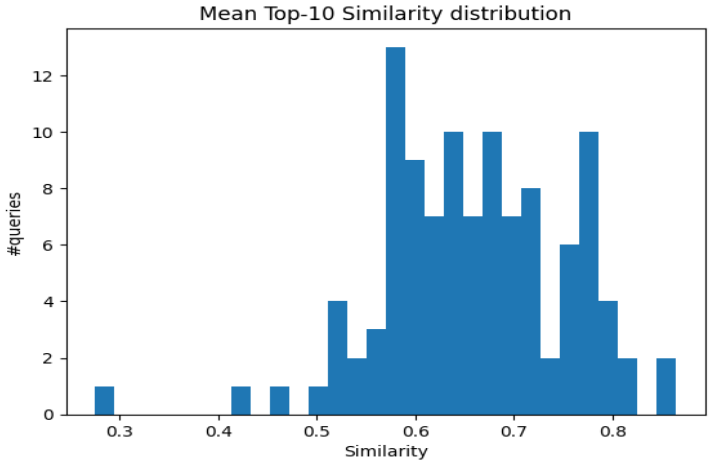


**Figure 3: Mean@Top -10 Distribution**

- As shown in Figure 4 ***The ranking drop (Top-1 minus Mean@10)*** was found to be only **0.074**, suggesting that the difference in quality between the best match and the rest is minimal. This is desirable in exploratory patent search tasks, where the user may wish to explore multiple high-relevance documents. The drop is small overall, with an average of 0.074 wherein most queries have a very small drop (<0.05). A smaller number of queries show a drop around 0.1–0.15, while very few queries have higher drops above 0.2 This parameter was chosen as part of the evaluation metric to measure not just accuracy of the top hit but stability of results across the top-k results. This ranking drop metric was part of the evaluation strategy to test ranking stability, and the low value directly supports my research question by showing that a lightweight, claim-aware similarity system can provide stable, interpretable, and exploratory search of patents for professional users.

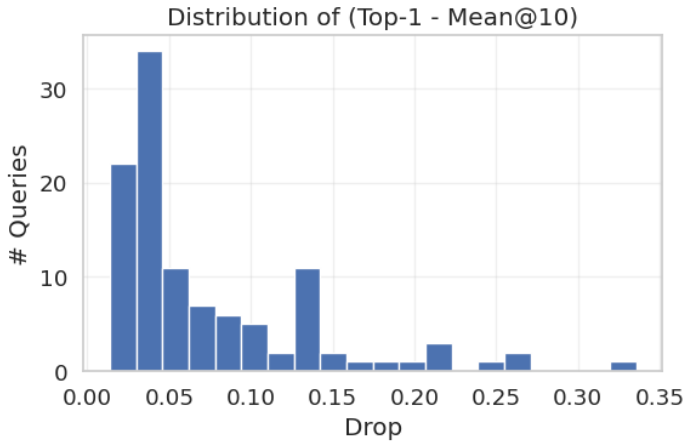


**Figure 4: Ranking Drop**

- Also, As shown in Figure 5, when measuring **threshold-based relevance**, an average of **7.3 out of 10 results exceeded a 0.6 similarity score**, and **3.36 exceeded 0.7**, with even **0.68 results per query crossing 0.8**, showing that a significant portion of each ranked list contains high-similarity results. These statistics provide strong evidence that the system retrieves not just one, but several useful results per query.

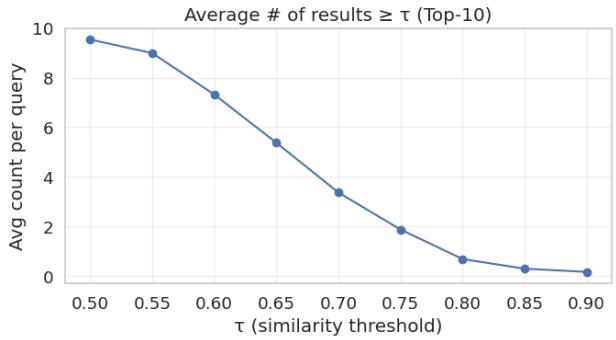


**Figure 5: Similarity threshold**

- Also, as per Figure 6 The **distribution characteristics** of similarity scores showed that most Top-1 results were concentrated between **0.68 to 0.76**, with very few below 0.60, as shown in the cumulative distribution function (CDF) and histogram plots. This suggests consistency in performance across the 110 diverse EV patents. (Use: *CDF + Histogram combo showing Top-1 distribution*)

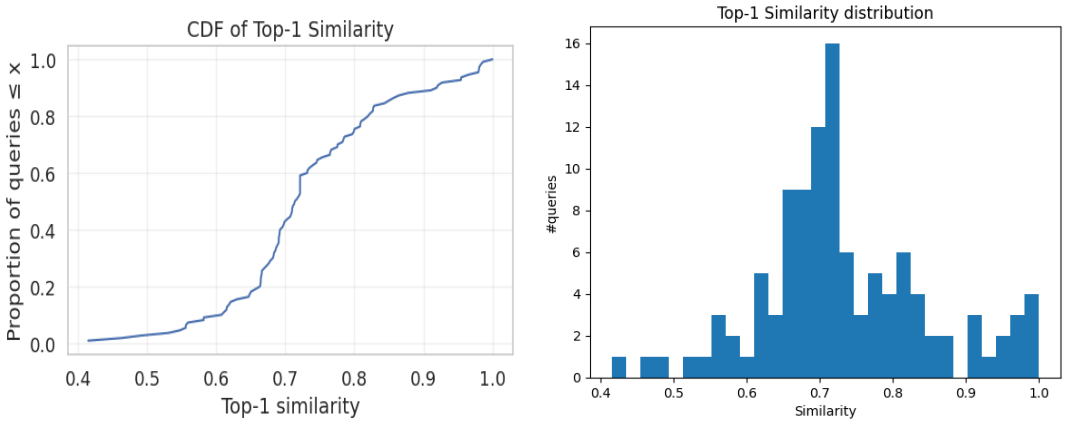


**Figure 6: Distribution and Cumulative Frequency of Top-1 Similarity Score**

- As per the Figure 7, Finally, the system was observed to perform better for **independent claims** compared to dependent claims, with average Top-1 similarity slightly higher for the former. This aligns with expectations, as independent claims tend to cover broader technical scope and thus yield more distinctive embeddings.

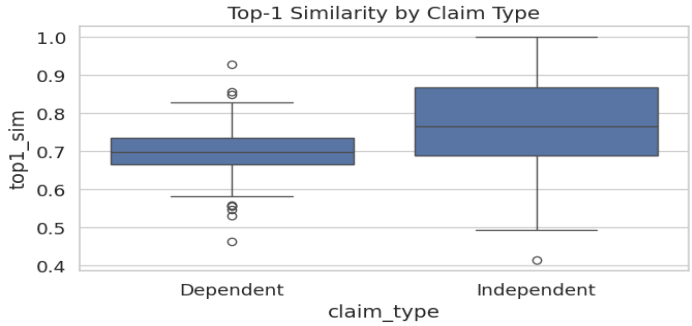


**Figure 7: Similarity by Claim Types**

Together, these insights validate the retrieval framework's capacity to deliver meaningful, stable, and semantically rich results for real-world EV patent search tasks. The system effectively avoids the binary constraints of traditional keyword-matching or citation-based techniques, favoring a nuanced similarity-based retrieval approach.

### 6.2 Experiment 2: Synthetic Query Evaluation

To further validate the robustness and semantic boundaries of the system, a second experiment was conducted using synthetically created patent claims. This test aimed to assess whether the retrieval system can distinguish between in-domain electric vehicle (EV)–related queries and queries that are either generic or irrelevant to the EV domain. Such an evaluation provides insight into how well the system can resist false positives and avoid returning misleading results for out-of-domain inputs.

Following synthetic Queries were generated in domains such as EV Battery, Braking System, Wireless Charging, Generic Tech, Totally irrelevant.

| Query Label | Synthetic Claim Description | Similarity Score |
|---|---|---|
| **EV Battery** | A battery management system for electric vehicles that optimizes charge cycles and prevents overheating. | 0.634 |
| **Braking System** | An autonomous vehicle braking system triggered by proximity sensors and predictive algorithms. | 0.566 |
| **Wireless Charging** | An electric vehicle system comprising a battery pack, regenerative braking, and wireless charging module. | 0.707 |
| **Generic Tech** | A mobile device that connects to cloud services and syncs user preferences. | ~0.45 |
| **Totally Irrelevant** | A method for growing tomatoes using AI-controlled sprinklers and UV lighting. | ~0.42 |

**Table 2: Details of Query and Score Experiment 2**

Each of these queries was processed through the same pipeline used for real EV claims — sentence embedding via MiniLM (msmarco-MiniLM-L6-v3) followed by similarity search using FAISS on the full 12k-patent corpus. The top 5 results per query were retrieved and analysed.

As shown, the top similarity scores decline as the query moves away from the EV domain. The system successfully ranked EV-specific claims higher, while generic or irrelevant queries yielded lower scores and unrelated patent results.

**Insights and Interpretation**

This experiment shows that the system is fairly capturing domain specific semantics. Wherein, High-relevance synthetic queries (like *Wireless Charging*) produced strong similarity scores (above 0.70), with retrieved patents clearly aligned with the EV space. Conversely, completely irrelevant queries such as the *Tomato AI sprinkler* example resulted in lower similarity scores (≈ 0.42) and no meaningful results from the EV dataset — confirming the system's resistance to false positives.

"The *Generic Tech* query, which consists of broadly technical but not explicitly EV-related terms, achieved moderately low similarity scores (~0.45). This suggests that the model does not overgeneralize to loosely related technological concepts, thereby maintaining domain specificity."

**Significance**

This experiment provides evidence that the model is not only accurate for known real-world queries but is also robust to unpredictable user inputs, such as edge-case, noisy, or out-of-domain queries. This robustness is critical in real-world deployment, where users may submit vague or exploratory claims. The model's ability to avoid irrelevant matches further enhances its reliability for professional use in patent intelligence and IP analytics.

### 6.3 Discussion

The evaluation results from both experiments provide compelling evidence that the developed patent similarity system meets its intended objectives: semantic relevance, domain specificity, and robustness across real-world and synthetic inputs.

From ***Experiment 1***, which tested the system using 110 real EV-domain patent queries not present in the main corpus, several strong patterns emerged. The Top-1 average similarity of 0.733 and mean@10 of 0.660 suggest not only that the top result is highly relevant but that the entire ranked set maintains strong semantic cohesion. This is a vital quality in intellectual property retrieval, where users often examine multiple documents for due diligence or novelty search. Furthermore, the drop between Top-1 and mean@10 was minimal (~0.074), confirming that the system does not merely spike at the first match but maintains relevance throughout the result list. This ranking consistency supports the core argument for using ranking-aware metrics over traditional binary relevance evaluations. Unlike binary assessments that merely confirm if a document is relevant, similarity scores offer a gradient of relevance, more aligned with legal and technical patent review workflows.

The system also showed domain awareness: over 70% of retrieved results had similarity scores above 0.6, reinforcing its capability to surface not just one, but multiple relevant entries. The claim-type breakdown further revealed that independent claims yielded higher Top-1 scores than dependent claims, indirectly validating the inclusion of claim weighting during preprocessing — a step inspired by legal practice but rarely applied in computational models.

According to Experiment 2 we have introduced different use case by including synthetic queries ranging from domain-relevant to entirely irrelevant. The system retrieved meaningful results for EV-focused queries like *Wireless Charging* and *EV Battery* (Top similarities ~0.71 and ~0.63, respectively), while effectively downranking unrelated ones such as *Tomato Farming using AI* (~0.41). The *Generic Tech* query yielded moderately low results (~0.45), suggesting the model avoids overgeneralization to loosely related technologies. This validates the domain-specific semantic filtering capability of the system and demonstrates its robustness in open-ended exploratory use cases.

Together, these results shows that the combination of transformer-based embeddings (MiniLM), structured claim weighting, and FAISS vector search developed a lightweight yet good performance system for patent similarity retrieval. Also, one of the Important benchmarks that this system achieves is semantic alignment without compromising on neither interpretability nor computational efficiency which is a balance often difficult to achieve in real-world information retrieval systems.

#### 6.3.1 User Interface (UI)

To support practical usability, the system includes a lightweight interface built using Stream lit, enabling users to input patent claims and retrieve the top-k semantically similar patents.

The UI supports both single-query and batch-mode search, with adjustable result count through a slider.

### a. Option 1: Single Query

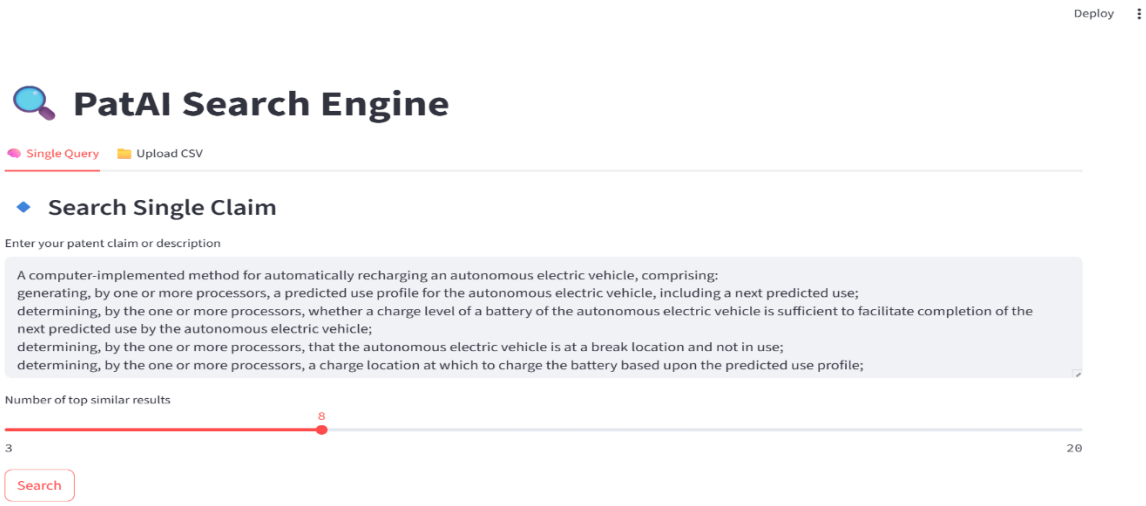


**Figure 8 shows the input interface where users can enter a claim and configure search settings.**

Upon submitting a query which is details of claims or key feature from patents or invention the system displays the most relevant patents along with their titles, publication numbers, and similarity scores, followed by a bar chart of top-k scores. And visibility of clear similarity scores gives confidence to person who is using the system regarding transparency and of the system.

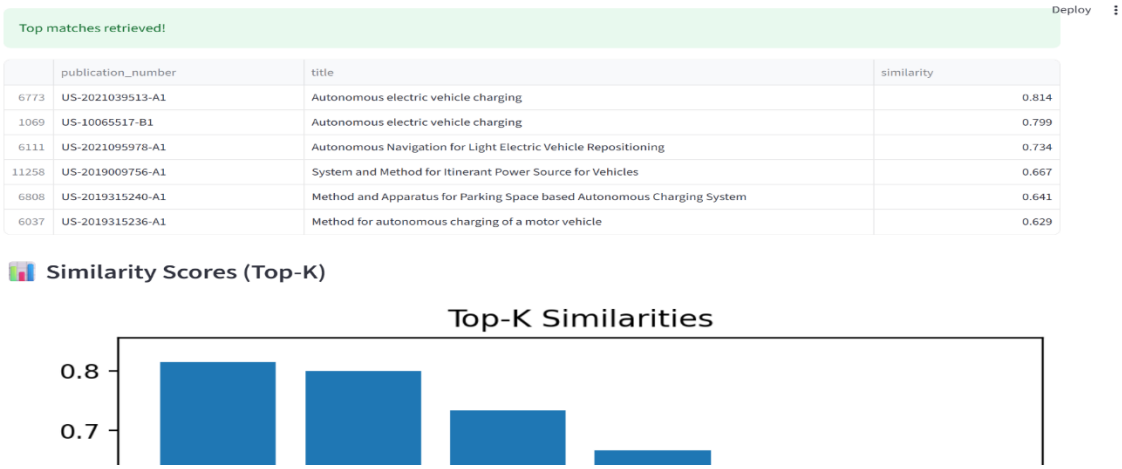


**Figure 9: Sample output with ranked results and score visualization.**

### b. Option 2: Batch Query

The system also provides a batch evaluation interface as per Fig & Fig, allowing users to upload a CSV containing multiple patent claims for simultaneous processing. This feature was instrumental in Experiment 1, enabling the evaluation of all 110 real EV queries (patent number and their subsequent details where claims are query text) at once. It outputs key metrics such as average top-1 similarity, mean top-k similarity, and a score distribution histogram (Figure 6.6), helping assess consistency across queries. This capability supports scalability and aligns with real-world use cases where bulk analysis is essential.

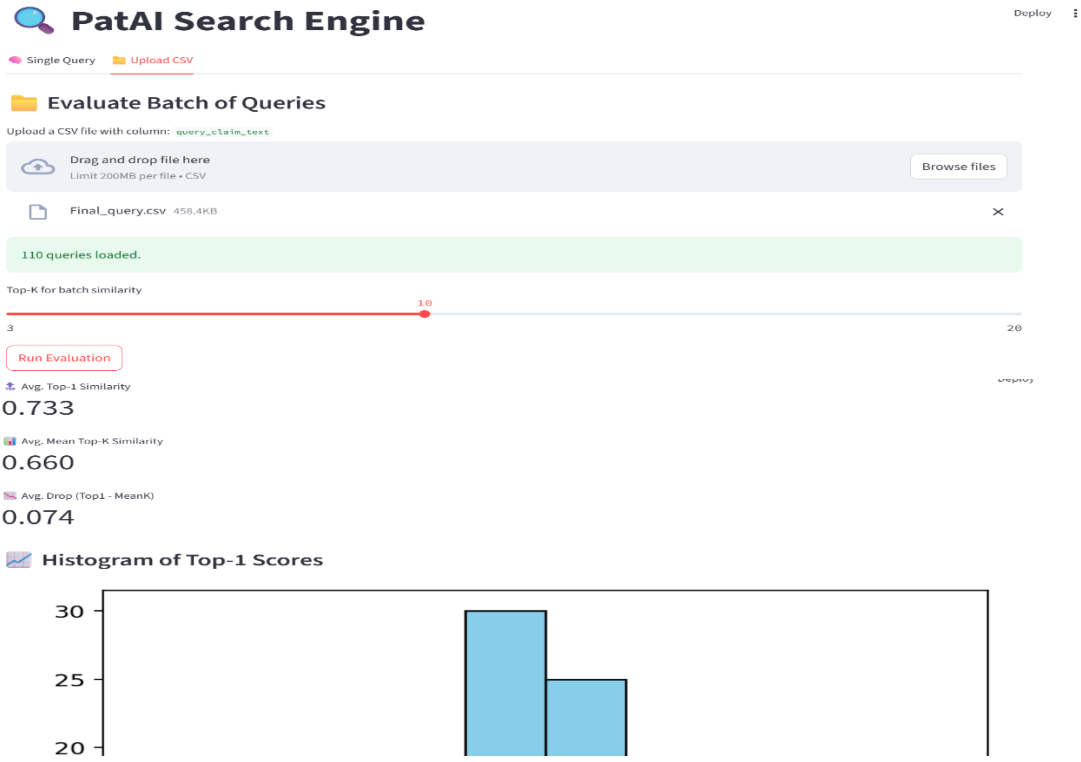


**Figure 10: Batch query output Screen**

For this project, I have focussed on designing a patent search system which is more efficient, transparent, and cost-effective. By introducing a AI based lightweight and claim focussed search system my project is offering a scalable solution that can be used by patent professionals, R&D teams, and corporate Intellectual Property departments. Following are business impacts of this project:

- Faster Prior Art Search & Reduced Legal Risk: This system targets this by ensuring relevant patents are not overlooked, lowering risks of litigation or invalidation.
- Cost Savings & Efficiency in Patent Analytics: As this system is lightweight in terms of computation when compared with already available solutions so can be used by smaller firms and thus its saves operation cost in running system like this.
- Transparency & Trust for Professionals: This project delivers a system that works by focussing on claims and it is interpretable so it promotes increase trust in results produced by patent search system.
- Scalability Across Domains: This project delivers a system that is simple yet efficient and can be used and implemented for not only Electric vehicle patents but patents from other domain such as medical, chemical, and telecom domains and many others that adds broader industry value.
- Decision Making for R& D: This project offers exploratory analysis meaning providing multiple patents/queries to identify white spaces, competitor activity, and investment opportunities when user want to search portfolio of patents.

# 7. Conclusion and Future Work

This chapter combine the findings of the research by summarizing the key outcomes, acknowledging the limitations, and explaining possible directions for future enhancement of the system.

## 7.1 Conclusion

This study developed a lightweight, domain-aware semantic similarity system for patents, specifically focused on the electric vehicle (EV) sector. By combining transformer-based embeddings (msmarco-MiniLM-L-6-v3), a claim-level weighting strategy, and FAISS-based retrieval, the system enables meaningful and scalable comparisons across patent claims. The model effectively retrieves relevant patents from a corpus of over 12,000 EV-related documents and delivers interpretable similarity scores to support IP analysis and exploration.

Through detailed evaluation using real-world and synthetic queries, the system demonstrated strong performance in identifying semantically close results, especially in structurally important independent claims. The high average similarity across the Top-10 results, as well as the system's ability to filter out unrelated queries, further validates its practical usefulness.

This work contributes both a functional tool and a methodological framework that balances accuracy, speed, and scalability that offers potential utility in R&D, patent monitoring, or litigation support.

### 7.2 Limitations

While the system performed well, certain limitations were observed, and they are

- **Limited diversity in queries**: Evaluation was primarily conducted using EV-domain queries. Use to other domains remains untested for both Main corpus and query dataset.
- **Unstructured metadata usage**: The system primarily leverages claim text without incorporating structured patent fields (e.g., CPC codes, inventors, citations), which could enhance performance.
- **Less Data**: The system is using only corpus of 12k US patent data as an initial setup. It can be scalable to exhaustive Electric vehicle patent dataset.

### 7.3 Future Work

The current system can be extended and enhanced in several directions:

- **Incorporate Retrieval-Augmented Generation (RAG)**: While the system currently acts as a "retriever," it can be extended into a RAG framework where retrieved results feed into a generative model for claim summarization, novelty detection, or report generation.
- **Add citation and CPC-based re-ranking**: Introducing additional metadata such as citations or classification codes can enable more nuanced scoring and cluster-aware retrieval.
- **Expand to multilingual or cross-jurisdictional patents**: Future work can support comparison between US, EP, and other international patents using language-agnostic embeddings.
- **Integrate user feedback for active learning**: Allowing users to mark relevant or irrelevant results could help retrain and personalize the model iteratively.
- **Deploy as a full web-based platform**: With the current Stream lit prototype in place, the system can be turned into a full-scale patent assistant tool with export options, analytics dashboard, and advanced filtering.
- **For other Domains**: The framework patent search system that I developed targeting Electric vehicle domain is lightweight and can be transitioned to other domains as

well such as medical chemical and other technological. The primary components embeddings, FAISS indexing, and claim-level search will be same however, primary dataset needs to be changed.

In conclusion, this study demonstrates the feasibility and value of using lightweight transformer models and semantic ranking techniques for patent similarity search in fast evolving and growing domains like electric vehicles. By adding claim-level understanding, semantic embeddings, and vector-based retrieval, the system provides a scalable and domain-aware solution. While the current implementation offers solid performance and interpretability, it also lays the groundwork for future improvements that can bring this system closer to real-world deployment in legal and IP analytics workflows.

# References


- Acikalin, U.U. and Kutlu, M. (2022) 'Patent Search Using Triplet Networks Based Fine-Tuned SciBERT', *arXiv*. Available at: https://doi.org/10.48550/arXiv.2207.11497 (Accessed: 11 September 2025).
- Ali, A., Khan, S., Rehman, F. and Zhang, L. (2024) 'Innovating Patent Retrieval: A Comprehensive Review of Techniques, Trends, and Challenges in Prior Art Searches', *Applied System Innovation*, 7(5), p. 91. Available at: https://doi.org/10.3390/asi7050091 (Accessed: 11 September 2025).
- Ascione, G.S. and Sterzi, V. (2024) 'A comparative analysis of embedding models for patent similarity', *arXiv*. Available at: https://doi.org/10.48550/arXiv.2403.16630 (Accessed: 11 September 2025).
- Bekamiri, H., Hain, D.S. and Jurowetzki, R. (2021) 'PatentSBERTa: A Deep NLP based Hybrid Model for Patent Distance and Classification using Augmented SBERT', *arXiv*. Available at: https://doi.org/10.48550/arXiv.2103.11933 (Accessed: 11 September 2025).
- Brean, D.H. (2018) 'Grading Patent Remedies: Dependent Claims and Relative Infringement'. Rochester, NY: *Social Science Research Network*. Available at: https://papers.ssrn.com/abstract=3252487 (Accessed: 11 September 2025).
- Douze, M., Johnson, J., Jégou, H. and Matuszewski, R. (2025) 'The Faiss library', *arXiv*. Available at: https://doi.org/10.48550/arXiv.2401.08281 (Accessed: 11 September 2025).
- Engineering at Meta (2017) 'Faiss: A library for efficient similarity search', *Engineering at Meta*, 29 March. Available at: https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/ (Accessed: 11 September 2025).
- Hain, D.S., Jurowetzki, R., Rakas, M. and Bekamiri, H. (2022) 'A text-embedding-based approach to measuring patent-to-patent technological similarity', *Technological Forecasting and Social Change*, 177, p. 121559. Available at: https://doi.org/10.1016/j.techfore.2022.121559 (Accessed: 11 September 2025).
- Helmers, L., Marx, A. and Schmoch, U. (2019) 'Automating the search for a patent's prior art with a full text similarity search', *PLOS ONE*, 14(3), p. e0212103. Available at: https://doi.org/10.1371/journal.pone.0212103 (Accessed: 11 September 2025).
- EPIC (no date) 'Industrial Policies and Innovation: Evidence from the Global Automobile Industry'. Available at: https://epic.uchicago.edu/research/industrial-policies-and-innovation-evidence-from-the-global-automobile-industry/ (Accessed: 11 September 2025).
- StartUs Insights (2025) 'EV Outlook 2025: Key Data & Innovations', *StartUs Insights*, 22 May. Available at: https://www.startus-insights.com/innovators-guide/ev-outlook/ (Accessed: 11 September 2025).
- Matthias, B., Heidari, G. and Hewel, C. (2024) 'Comparing Complex Concepts with Transformers', in *Proceedings of the 1st Workshop on Legal and Patent Text Analytics (LPText 2024)*. CEUR-WS.org. (Accessed: 11 September 2025).
- SBERT (no date) 'MSMARCO Models — Sentence Transformers documentation'. Available at: https://www.sbert.net/docs/pretrained-models/msmarco-v3.html (Accessed: 11 September 2025).

- Vowinckel, K. and Hähnke, V.D. (2023) 'SEARCHFORMER: Semantic patent embeddings by siamese transformers for prior art search', *World Patent Information*, 73, p. 102192. Available at: https://doi.org/10.1016/j.wpi.2023.102192 (Accessed: 11 September 2025).
- Wang, W., Bao, H., Dong, L. and Wei, F. (2020) 'MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers', *arXiv*. Available at: https://doi.org/10.48550/arXiv.2002.10957 (Accessed: 11 September 2025).
- Yoo, Y., Li, H., Xu, Q. and Cao, L. (2023) 'A Novel Patent Similarity Measurement Methodology: Semantic Distance and Technological Distance', *arXiv*. Available at: https://doi.org/10.48550/arXiv.2303.16767 (Accessed: 11 September 2025).
- Yoo, Y., Xu, Q. and Cao, L. (2025) 'PatentMind: A Multi-Aspect Reasoning Graph for Patent Similarity Evaluation', *arXiv*. Available at: https://doi.org/10.48550/arXiv.2505.19347 (Accessed: 11 September 2025).
- Zhu, M., Li, Q., Zhang, Y. and Chen, W. (2023) 'A multi-dimensional fusion strategy similarity measure method for patent application technology disclosure document', *PLOS ONE*, 18(10), p. e0293091. Available at: https://doi.org/10.1371/journal.pone.0293091 (Accessed: 11 September 2025).