Document Relevance Filtering by Natural Language Processing and Machine Learning: A Multidisciplinary Case Study of Patents

Raj Bridgelall *

Posted Date: 24 January 2025

doi: 10.20944/preprints202501.1853.v1

Keywords: document search; supervised machine learning; unsupervised machine learning; natural language processing; latent Dirichlet allocation; non-negative matrix factorization; manifold learning; t-distributed stochastic neighbor embedding; term co-occurrence networks

 doi:10.20944/preprints202501.1853.v1

Article

Document Relevance Filtering by Natural Language Processing and Machine Learning: A Multidisciplinary Case Study of Patents

Raj Bridgelall

Department of Transportation and Supply Chain, College of Business, North Dakota State University, P.O. Box 6050, Fargo, ND 58108-6050, USA; raj@bridgelall.com or raj.bridglall@ndsu.edu

Abstract: The exponential growth of patent datasets poses a significant challenge in filtering relevant documents for research and innovation. Traditional semantic search methods based on keywords often fail to capture the complexity and variability of multidisciplinary terminology, leading to inefficiencies. This study addresses the problem by systematically evaluating supervised and unsupervised machine learning (ML) techniques for document relevance filtering across five technology domains: solid-state batteries, electric vehicle chargers, connected vehicles, electric vertical takeoff and landing aircraft, and light detecting and ranging (LiDAR) sensors. The contributions include benchmarking the performance of 10 classical models. These models include extreme gradient boosting, random forest, and support vector machines; a deep artificial neural network; and three natural language processing methods: latent Dirichlet allocation, non-negative matrix factorization, and k-means clustering of a manifold-learned reduced feature dimension. Applying these methods to more than 4,200 patents filtered from a database of 9.6 million patents revealed that most supervised ML models outperform the unsupervised methods. An average of seven supervised ML models achieved significantly higher precision, recall, and F1-scores across all technology domains, while unsupervised methods show variability depending on domain characteristics. These results offer a practical framework for optimizing document relevance filtering, enabling researchers and practitioners to efficiently manage large datasets and enhance innovation.

Keywords: document search; supervised machine learning; unsupervised machine learning; natural language processing; latent Dirichlet allocation; non-negative matrix factorization; manifold learning; t-distributed stochastic neighbor embedding; term co-occurrence networks

1. Introduction

In an era of rapid technological innovation, effectively filtering relevant documents from large databases is essential for research, planning, and informed decision-making. Traditional semantic searches using keywords, while useful, often return a mix of relevant and irrelevant documents [1]. This is because different terms can describe similar concepts, or identical terms can have multiple meanings across different technical fields. Patents in particular employ broad, non-standard terminology to maximize claim scope, making exact keyword matches unreliable [2].

The **goal** of this study is to develop and evaluate natural language processing (NLP) and machine learning (ML) techniques that can efficiently and accurately filter documents relevant to a specific technology domain. This study applied four sets of methods for comparison. They comprised two NLP techniques: topic modeling using latent Dirichlet allocation (LDA) and non-negative matrix factorization (NMF); an unsupervised ML (UML) method that clustered features extracted from keyword frequency and positional occurrence, dimensionally reduced by t-distributed stochastic neighbor embedding (t-SNE), supervised ML (SML) methods based on statistical models like gradient boosting, random forest, and support vector machines; and a deep learning model based on artificial neural networks (D-ANNs).

 doi:10.20944/preprints202501.1853.v1

This study applied the methods to five technology domains spanning innovations in transportation systems: solid-state batteries (SSBs), electric vehicle chargers (EVCs), connected vehicles (CVs), electric vertical takeoff and landing (eVTOL) aircraft, and light detecting and ranging (LiDAR) sensors enabling autonomous vehicles. The **contribution** of this research is to benchmark the performance of the four sets of methods proposed to optimize document relevance filtering in these complex and dynamic information domains, enabling ongoing research and technology forecasting.

The rest of this paper proceeds as follows: Section 2 reviews the literature on methods of document relevance filtering. Section 3 explains the data source, visualizes relevant and irrelevant topics retrieved from each technology domain using a semantic search, and describes the four sets of methods proposed for document relevance filtering. Section 4 benchmarks the performance of the four sets of methods. Section 5 discusses the results, implications, and limitations of the study. Section 6 concludes the research and suggests future work.

2. Literature Review

Document relevance filtering has long been a challenge due to the sheer volume of unstructured text data and the ambiguity in terminology. Traditional approaches, such as semantic searches using keywords, rely on exact or partial matches and often return a mix of relevant and irrelevant results, particularly in multidisciplinary fields [3]. Patent documents in particular frequently use broad, non-standardized language to maximize claim scope, further complicating relevance filtering [4]. In a recent review of 78 articles, Ali et al. (2024) point out many of the gaps remaining in the field of document relevance filtering and concluded that despite advancements so far, the technical nature of documents such as those employing patent lingo increases the difficulty of retrieving high-relevance documents [2].

Recent advances in NLP and ML have introduced more sophisticated methods for addressing these challenges [5]. NLP techniques have gained traction in processing patent text [6], with topic modeling techniques such as LDA and NMF providing a probabilistic framework to uncover latent themes in text corpora and serve as a method of topic filtering [7]. However, limitations in their ability to align topics with document relevance highlight the need for improved feature engineering and domain-specific adjustments [8].

Researchers have widely adopted SML techniques like gradient boosting (GB), random forests (RFs), and support vector machines (SVMs) for various applications of text processing [9]. However, only a few studies have recently adopted SML methods for patent retrieval and relevance classification [10]. D-ANNs have also emerged as powerful tools to mine high-dimensional text data, leveraging their ability to learn nonlinear and hierarchical feature representations [11]. However, D-ANNs have high computational demands [12].

Data visualization methods like word clouds and term co-occurrence networks facilitate exploratory analysis, enabling researchers to identify patterns and relationships in large datasets [13]. While promising, their effectiveness heavily depends on feature selection and the quality of input data [14]. The current study proposed applying t-SNE, a dimensionality reduction technique, coupled with clustering algorithms like k-means, to help visualize and segregate text data.

The application of ML techniques in patent analysis remains nascent but increasingly critical [15]. Patents are a rich source of technological insights, and efficiently filtering relevant documents can significantly enhance innovation forecasting and strategic planning. Yet, there is a lack of recent studies focused on patent document relevance filtering as evidenced by the few documents retrieved in this literature review. The studies that focused on transportation technologies, such as solid-state batteries and autonomous vehicle systems, highlight the need for tailored ML solutions that account for domain-specific nuances and evolving terminologies. By building on this foundation, the current study seeks to bridge gaps in the existing literature and provide a benchmark for relevance filtering in multidisciplinary patent datasets.

 doi:10.20944/preprints202501.1853.v1

3. Methodology

Error! Reference source not found. illustrates the methodological workflow of the study. The workflow begins by defining the five technology domains for the case study and defining the search keywords for the selected database.

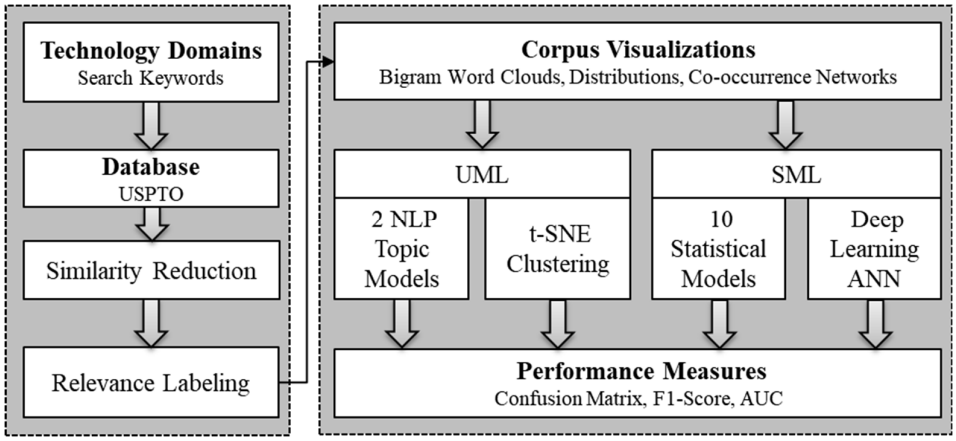

Figure 1. The methodological workflow.

Many patent documents are similar because they apply small changes to their claim set, with much of the earlier description of the technology being mostly unchanged. Hence, the workflow applied a cosine similarity measure to the vectorized text data and removed documents that were more than 90% similar. The author then read and labeled each document for relevance. The subsections that follow further describe the workflow.

3.1. Database

This study selected the United States Patent and Trademark Office (USPTO) patent summary database due to its critical role as a repository of global innovation. As the world's largest and most important commercial market, the United States attracts a high volume of patent filings from both domestic and international entities seeking to secure intellectual property rights [16]. For instance, more than half of USPTO filings come from international assignees, who often prioritize the USPTO for their initial or subsequent filings to establish a strong market position and gain legal protection in the U.S. [17]. This international dimension makes the USPTO database a rich source of diverse and high-quality patent documents, representing innovative advancements across the selected technology domains. **Error! Reference source not found.** summarizes the search keywords and provides examples of relevant and irrelevant topics retrieved for each technology domain. Some of the keywords utilize wildcard characters like "*" and "?" to instruct the search engine to retrieve documents containing word variations such as their plural form and those containing keywords that ignore word separators like dashes.

 doi:10.20944/preprints202501.1853.v1

Table 1. Search keywords and examples of relevant and irrelevant topics.

Domain	Keywords	Examples of Relevant Topics	Examples of Irrelevant Topics
SSB	*solid?state?batter*, vehicle*	Challenges and solutions to SSB design including resolving interfacial resistance, ionic conductivity, charge cycling stability, and energy density through new architectures, electrodes, electrolytes, and manufacturing strategies.	Designs focused on battery chargers, circuitry, parking optimization to access chargers, connectivity to chargers, cloud services to evaluate charging levels, overcharge prevention, charge scheduling strategies, and optimized route planning to access charger networks.
EVC	electric vehicle, charger, batter*	Challenges and solutions to charge scheduling, charge station management, charging device controller, charging safety, fast charging, fuel integration, multiple vehicle charging, on-board charger designs, power transfer efficiency, charge state monitoring, and temperature management.	Charge pricing strategies, marketing promotions, smart grid and energy storage conversion systems, locating charging facilities, robotic chargers, vehicle sharing management, hybrid-electric vehicles, solar chargers, vehicle-to-grid energy management, electric motor design, motor controller designs, and regenerative power management.
CV	"connected vehicle"	Wireless communications methods, driving safety, navigation, condition monitoring, traffic signal adaptation, smart parking, multi-vehicle networking, vehicular information management, cybersecurity, and vehicle computing resource optimization.	Traffic demand modeling, cloud service hosting, customized vehicle settings, road network mapping, driver distraction detection, garage door control, towed vehicles, trailer controls, overhead line system, in-vehicle networks, edge-computing, wired base stations, and vehicle tethered apparatus or trailers.
eVTOL	aircraft, vertical, electric	Adaptable systems like rotors and flaps, control enhancement, drag reduction, energy management, propulsion efficiency, safety enhancement, thermal management, and transition efficiency.	Refrigerated cargo bays, user interfaces, stress sensors, aircraft simulators, combustion engine cooling, airbag integration, image capturing, remote sensing, display devices, remote control, and aircraft headlamps.
LiDAR	lidar, vehicle, distance, cost	Beam generation, beam steering, cost or size reduction, field-of-view enhancement, and signal quality management.	Road signage mapping, camera integration, vehicle control, left-behind object recognition, drone-based remote sensing, and vehicle remote control.

 doi:10.20944/preprints202501.1853.v1

Error! Reference source not found. summarizes the sizes of the corpora searched, the number of matching documents, the number of documents removed after filtering for similarity, and the number of relevant documents identified after labeling. In particular, searching the corpora of 9.6 million documents spanning 2018 to 2024 resulted in 4,214 containing matching keywords and, of those, 1,201 covered relevant topics. Only an average of 29% of the documents with matching keywords were relevant, quantifying the inefficiency and confirming the low effectiveness of keyword-based searches. The SSB case resulted in the largest proportion of documents with matching keywords being relevant at 87% and eVTOL with the lowest at 11%, highlighting the variability of keyword-based search effectiveness across technology domains.

Table 2. Search results from the patent database.

Parameter	SSB	EVC	CV	eVTOL	LiDAR	Total
Corpus Size	2,160,576	1,637,725	1,637,725	1,995,672	2,239,447	9,671,145
Keyword Filter	288	1,233	622	1,472	867	4,482
Similarity Filter	281	1,184	600	1,413	736	4,214
Labeled Relevant	245	390	220	158	188	1,201
RI Ratio	0.87	0.33	0.37	0.11	0.26	

3.2. Visualizing Data

Assessing the quality of extracted textual data and evaluating the differences between relevant and irrelevant documents by reading a large corpus is laborious and inefficient. Therefore, the present research proposed a combination of bigram word clouds, bigram distributions, and term co-occurrence networks to quickly visualize, cross correlate, and assess topics across the corpus of each technology domain. These methods were effective in helping to identify distinguishing features between relevant and irrelevant documents at a glance. However, before generating these visualizations, text cleaning and lemmatization was necessary to enhance information content (signal) while reducing the number of non-contextual words (noise). That is, text cleaning boosted the signal-to-noise (SNR) ratio of documents by removing short words, digits, punctuation symbols, and common stop words such as “should,” “have,” and “they” that lack context. Lemmatization standardized words to their root forms, which eliminated redundancies from word variations and improved processing efficiency.

The **bigram word cloud** visualizes the most frequently occurring adjacent word pairs (bigrams), with font sizes adjusted to represent their relative frequencies. Hence, word clouds provide a qualitative overview of prominent terms to assess their topic associations. The **bigram distribution** (implemented as a histogram) complements the bigram word cloud by quantitatively conveying the relative importance of the most common bigrams. The **term co-occurrence network** highlights the relationships between key terms in the corpus, visually conveying their frequencies and connection strengths in document pairs. The term co-occurrence network also reveals clusters of related documents, offering insight into the structure and subtopics within the corpus. Together, these visualizations enable a more intuitive and accessible exploration of topic differences in the corpus. The subsections that follow demonstrate how these tools provide a comparative view of relevant and irrelevant document sets for each technology domain.

3.2.1. SSB

Error! Reference source not found. illustrates bigram visualizations for relevant and irrelevant documents within the SSB corpus. **Error! Reference source not found.**a shows the bigram word cloud and frequency distribution for the irrelevant document set. Common terms such as “electronic device,” “energy storage,” and “neural network” dominate the visualizations, reflecting adjacent but unrelated technological topics. **Error! Reference source not found.**b depicts the relevant document set, highlighting SSB-specific terminology such as “active material,” “positive electrode,” “negative electrode,” and “electrolyte layer.” These terms are core components of SSB development and are

 doi:10.20944/preprints202501.1853.v1

consistently more prominent in both the word cloud and frequency distribution. In this domain, the irrelevant documents exhibit broad and less focused topics, whereas the relevant documents emphasize the core concepts and terminologies critical to SSB innovation.

Error! Reference source not found. provides a complementary visualization of the SSB corpus by using the term co-occurrence networks for irrelevant and relevant document sets. **Error! Reference source not found.**a represent the irrelevant documents, where broad terms such as "vehicle," "electronic device," and "terminal," dominate the network, reflecting connections across a range of adjacent topics. The clustering in this network shows weaker thematic cohesion, with terms like "neural network" and "power storage device" loosely associated, indicating a lack of focus on SSB-specific terminology.

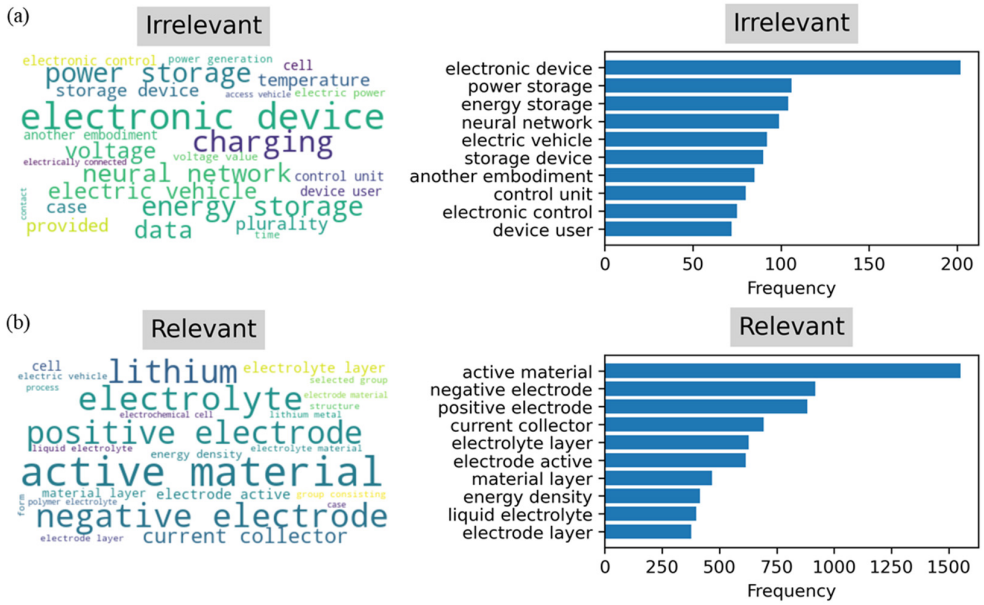

Figure 2. SSB top bigrams in a) irrelevant and b) relevant documents.

Error! Reference source not found.b, on the other hand, visualizes the relevant document set, where key terms such as "solid-state battery," "negative electrode," and "solid-state electrolyte" emerge as central nodes. The network shows strong interconnections among these core SSB terms, including related components like "cathode," "anode," "sulfide," and "electrolyte layer." These clusters reveal the thematic coherence of the relevant documents, reflecting the technical focus on SSB research and development. Hence, the term co-occurrence networks effectively highlight the structural differences between the irrelevant and relevant document sets. The tightly knit connections in the relevant network provide a visual representation of topic specificity, while the dispersed and less focused clusters in the irrelevant network highlight the challenge of identifying relevance in a large corpus.

 doi:10.20944/preprints202501.1853.v1

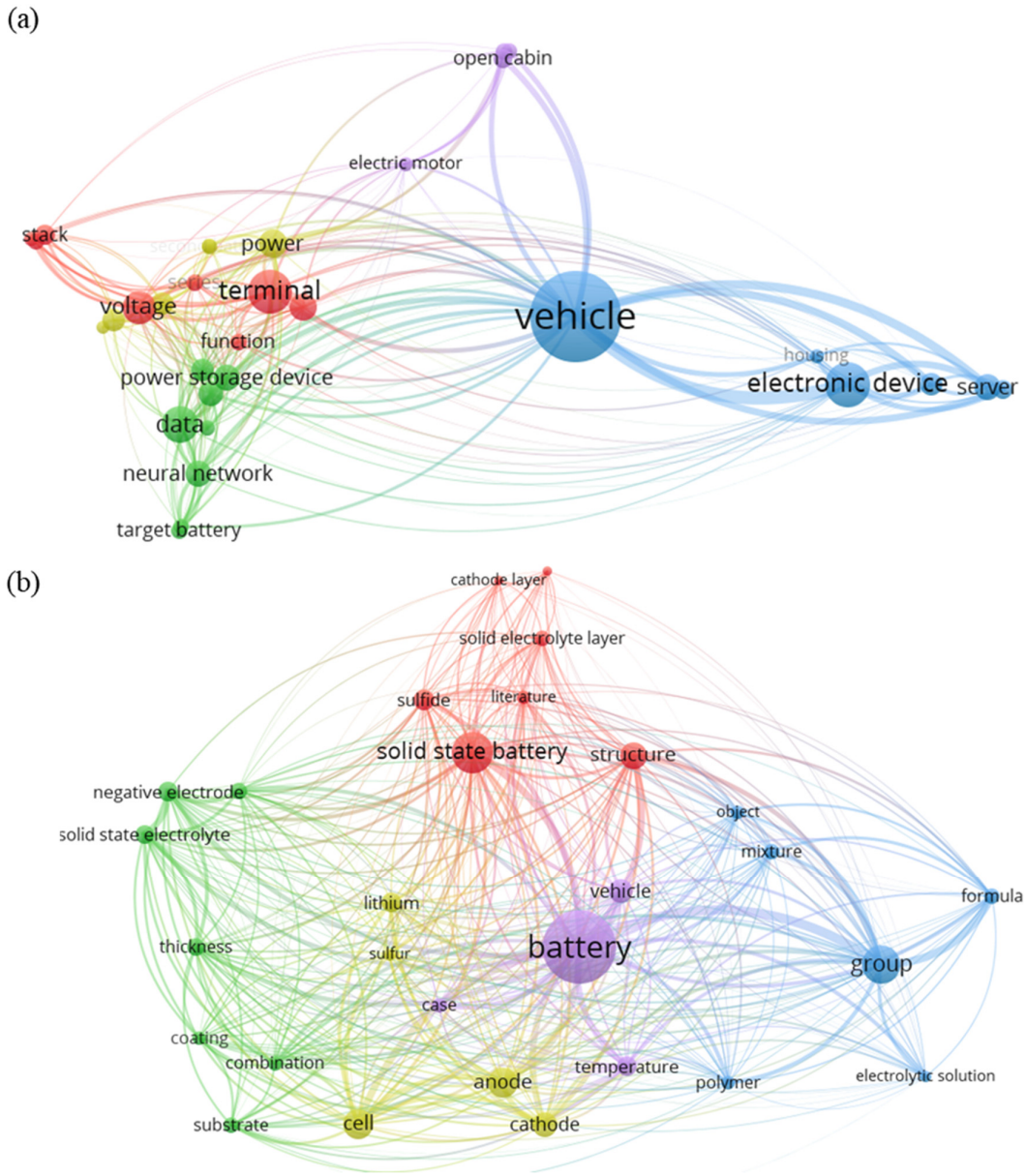

Figure 3. SSB term co-occurrence for a) irrelevant and b) relevant documents.

3.2.2. EVC

Error! Reference source not found. presents the bigram word cloud and frequency distribution for the EVC corpus, comparing irrelevant and relevant document sets. **Error! Reference source not found.**a showcases the irrelevant documents, where prominent terms such as "wireless power," "battery pack," and "energy storage" dominate the word cloud and frequency distribution. These terms, while related to broader energy and charging contexts, lack specificity to electric vehicle charger designs. **Error! Reference source not found.**b focuses on the relevant documents, with terms like "charger," "electric," "battery charging," and "energy storage" emerging as central themes in both the word cloud and frequency distribution. These bigrams emphasize the technical focus on EVC-related topics, such as power supply systems, charging mechanisms, and electric vehicle infrastructure.

Error! Reference source not found. illustrates the term co-occurrence networks for irrelevant and relevant document sets in the EVC corpus. **Error! Reference source not found.**a represents the irrelevant documents, where terms such as "system," "power," and "device" dominate as central

 doi:10.20944/preprints202501.1853.v1

nodes. The network features loosely connected clusters related to general topics like wireless power and vehicle components, reflecting a lack of focus on EVC designs.

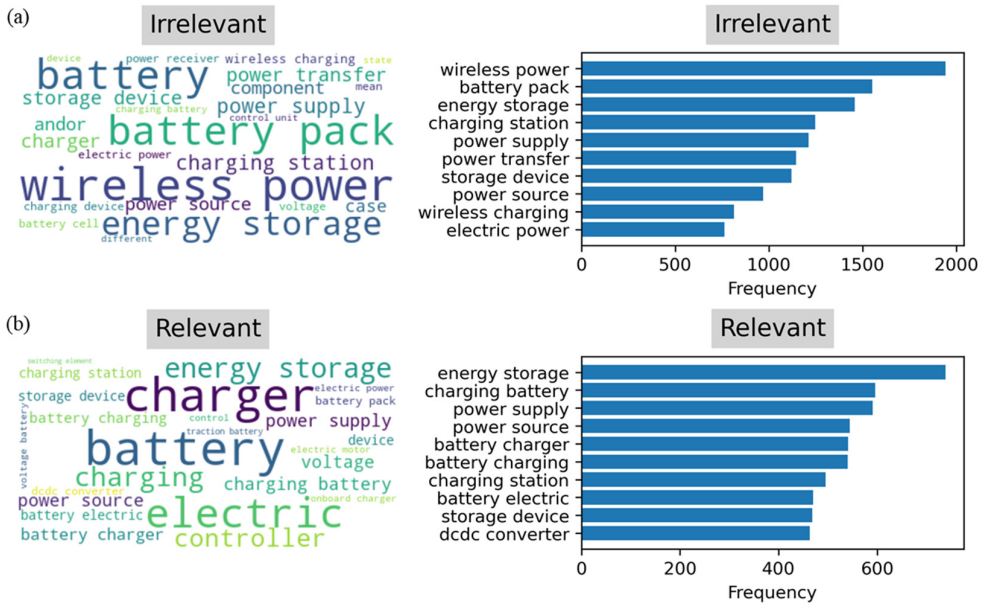

Figure 4. EVC top bigrams in a) irrelevant and b) relevant documents.

Error! Reference source not found.b visualizes the relevant document set, showcasing a more cohesive network centered around EVC design terms such as “battery,” “electric vehicle,” and “voltage.” This network reveals stronger interconnections among key EVC design concepts, including “converter,” “circuit,” and “dc-dc converter,” emphasizing the technical focus on designing charging mechanisms and implementing electric vehicle infrastructure.

 doi:10.20944/preprints202501.1853.v1

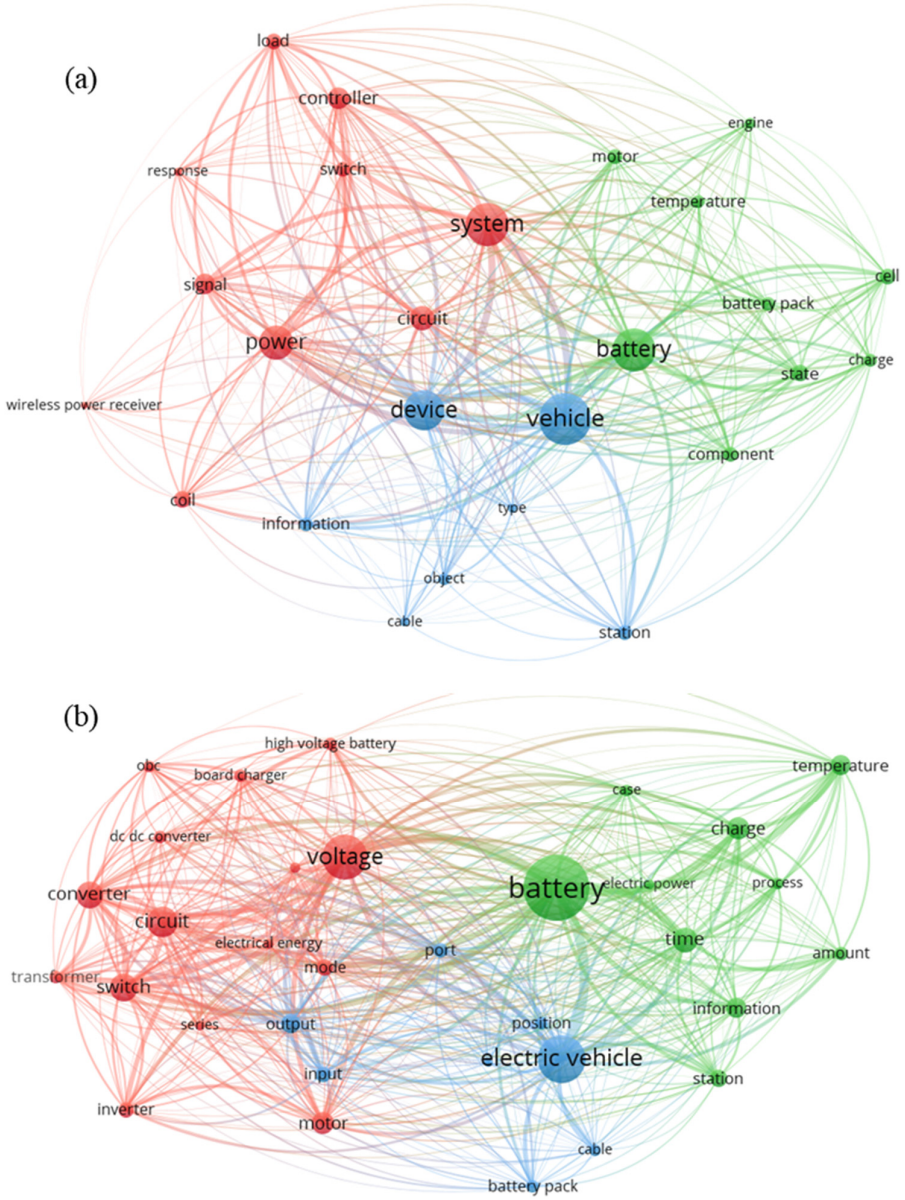

Figure 5. EVC term co-occurrence for a) irrelevant and b) relevant documents.

3.2.3. CV

Error! Reference source not found. compares the bigram visualizations for irrelevant and relevant document sets within the CV corpus. **Error! Reference source not found.**a presents the irrelevant documents, where terms like "wireless service," "base station," and "operation" are prevalent in both the word cloud and frequency distribution. These terms reflect general communication and networking concepts that are tangentially related to CV technologies but do not focus on their design elements. **Error! Reference source not found.**b depicts the relevant document set, highlighting bigrams such as "micro cloud," "perform action," and "executed processor." These terms closely align with the functional elements of CV technology, including the integration of cloud computing, autonomous functionality, and advanced processing systems. The word cloud emphasizes the prominence of these terms, while the frequency distribution quantifies their relative importance within the relevant corpus.

 doi:10.20944/preprints202501.1853.v1

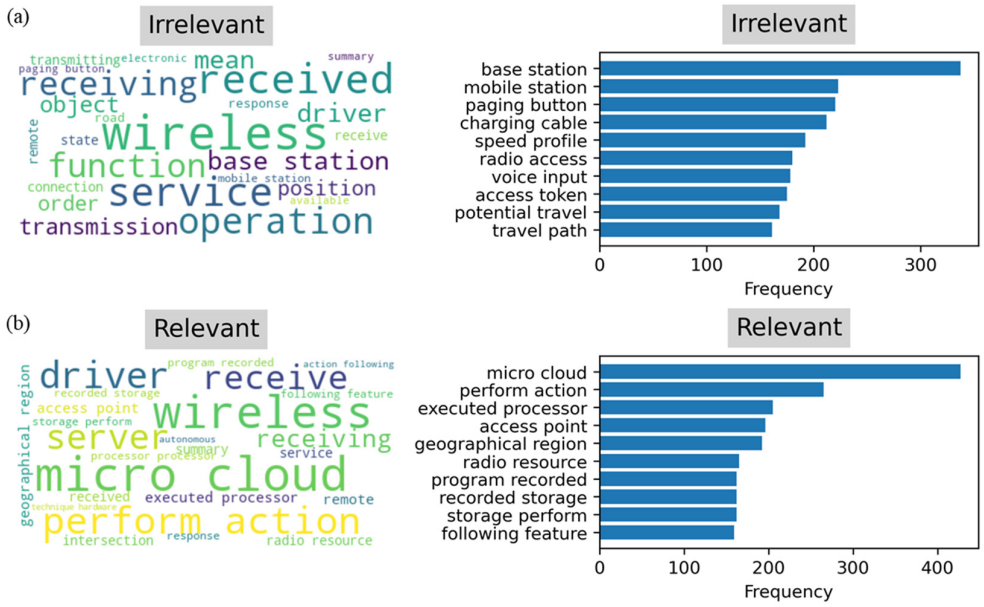

Figure 6. CV top bigrams in a) irrelevant and b) relevant documents.

Error! Reference source not found. illustrates the term co-occurrence networks for irrelevant and relevant document sets in the CV corpus, highlighting the relationships and clusters of key terms. **Error! Reference source not found.**a visualizes the irrelevant document set, where terms such as "vehicle," "system," and "device" dominate as central nodes. The network shows dispersed clusters related to broad concepts like "communication," "signal," and "mobile device," reflecting general connectivity and device-oriented themes that lack specificity to CV technologies.

Error! Reference source not found.b represents the relevant document set, where the network exhibits tighter clusters and stronger connections among terms related to CVs. Central terms like "vehicle," "data," and "computer" dominate the visualization, with additional focus on CV design or operational themes such as "sensor data," "micro cloud," and "ADAS system." These clusters highlight the integration of advanced data processing, cloud computing, and autonomous systems within CV technologies.

 doi:10.20944/preprints202501.1853.v1

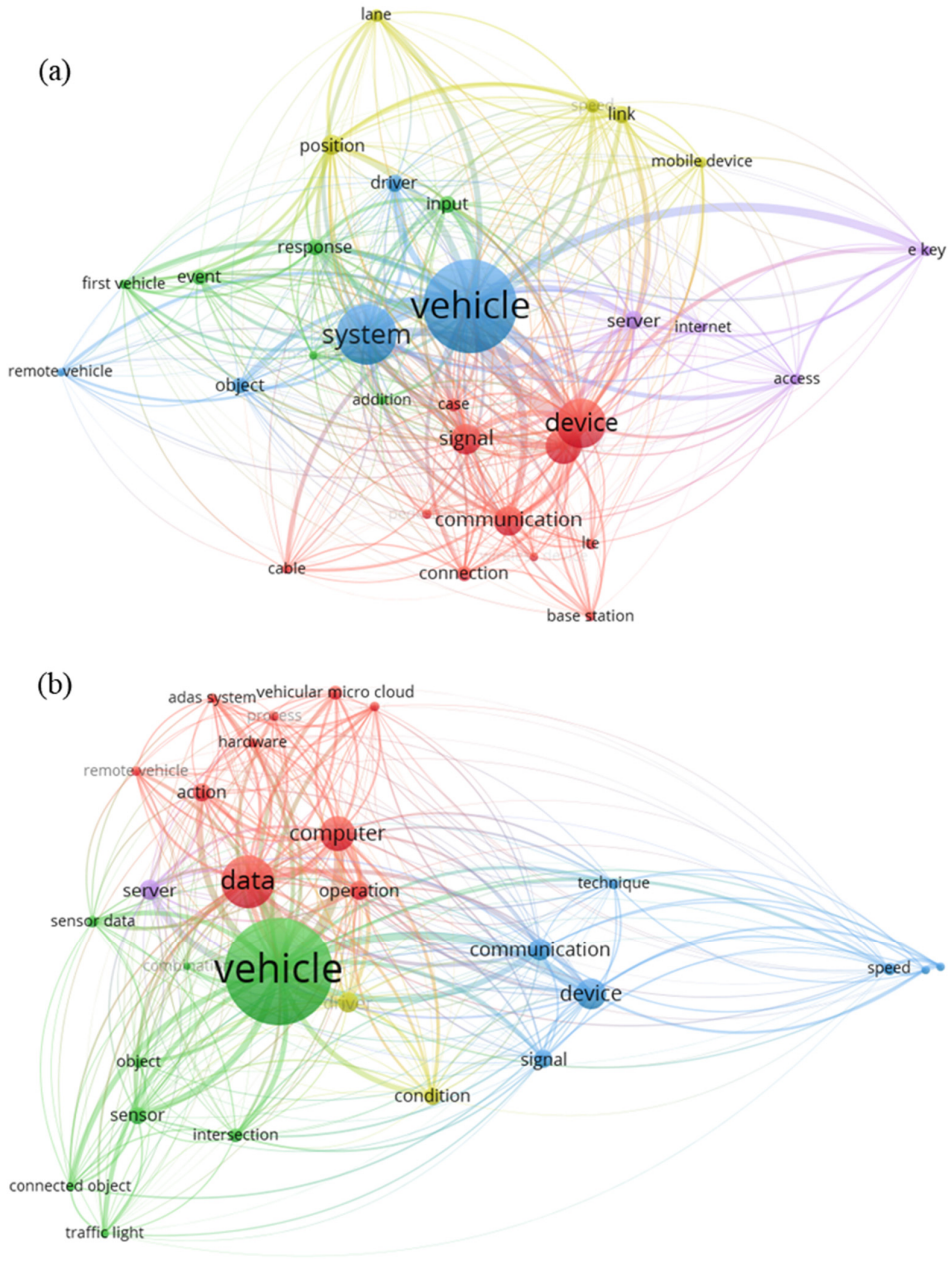

Figure 7. CV term co-occurrence for a) irrelevant and b) relevant documents.

3.2.4. eVTOL

Error! Reference source not found. compares the bigram word clouds and frequency distributions for irrelevant and relevant document sets in the eVTOL corpus. **Error! Reference source not found.**a illustrates the irrelevant documents, where terms such as "control component," "connection element," and "optical sensor" dominate. These terms reflect a broad range of general engineering and component-related topics but lack specificity to eVTOL technology. Conversely, **Error! Reference source not found.**b focuses on the relevant documents, where bigrams like "rotor assembly," "propulsion assembly," and "electrical power" emerge as key themes. These terms emphasize critical eVTOL design concepts, such as propulsion systems, power distribution, and aerodynamic control. The frequency distribution further highlights the importance of these terms, emphasizing their prevalence in the relevant document set.

 doi:10.20944/preprints202501.1853.v1

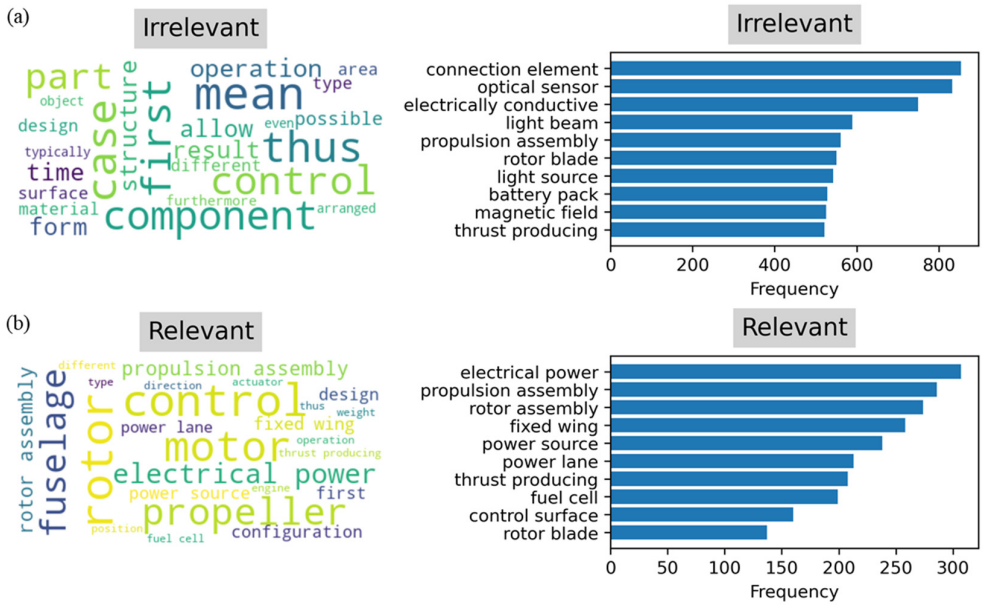

Figure 8. eVTOL top bigrams in a) irrelevant and b) relevant documents.

Error! Reference source not found. illustrates the term co-occurrence networks for irrelevant and relevant document sets within the eVTOL corpus, highlighting the relationships and thematic structures of key terms. **Error! Reference source not found.**a represents the irrelevant document set, where terms such as “device,” “element,” and “aircraft” dominate the network. The clusters are broad and loosely connected, encompassing general topics like “sensor,” “component,” and “field,” which are not specific to eVTOL technology development. These dispersed connections reflect the lack of thematic focus within the irrelevant set. **Error! Reference source not found.**b visualizes the relevant document set, showcasing a more cohesive network centered around terms directly associated with the design of eVTOL systems. Key nodes such as “aircraft,” “rotor,” and “wing” dominate, with tightly connected clusters focusing on critical concepts like “propulsion system,” “vertical takeoff,” and “flight control.” The network reflects the specific design challenges and solutions within the eVTOL domain, emphasizing interrelations among propulsion, aerodynamics, and power systems.

 doi:10.20944/preprints202501.1853.v1

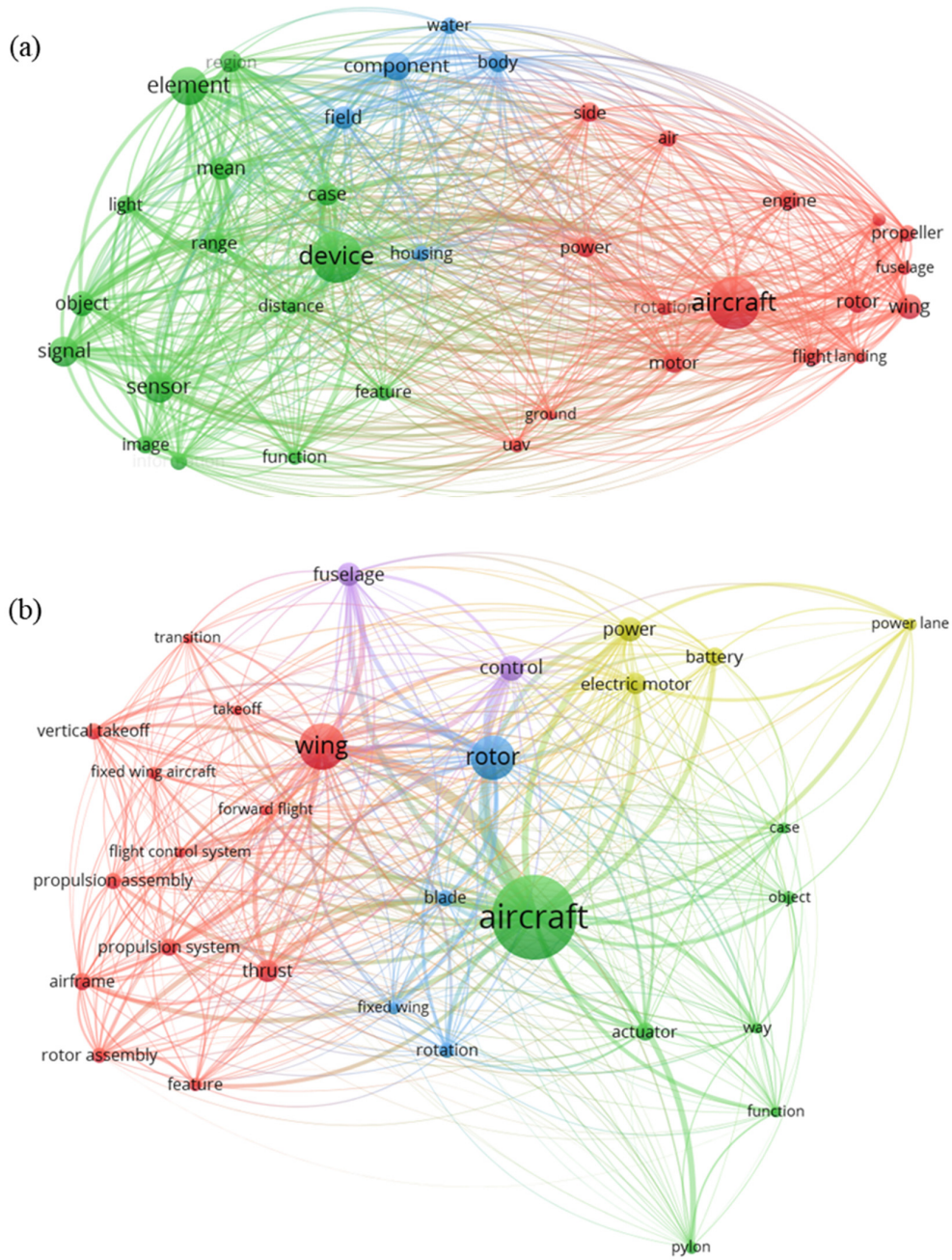

Figure 9. eVTOL term co-occurrence for a) irrelevant and b) relevant documents.

3.2.5. LiDAR

Error! Reference source not found. compares the bigram visualizations for irrelevant and relevant document sets within the LiDAR corpus. **Error! Reference source not found.**a depicts the irrelevant documents, where terms such as “vehicle,” “point cloud,” and “sensor data” dominate. These terms suggest a broad focus on general automotive and sensor systems, which are tangentially related to LiDAR-specific technologies. The word cloud and frequency distribution highlight the prevalence of generic terms like “host vehicle” and “control system,” reflecting a lack of focus on LiDAR designs. **Error! Reference source not found.**b presents the relevant document set, where bigrams such as “field view,” “laser beam,” and “optical system” are prominent. These terms directly align with the design of LiDAR systems, including optical and laser technologies central to its operation. The frequency distribution quantifies the dominance of design terms related to LiDAR, reinforcing their importance in the relevant corpus.

 doi:10.20944/preprints202501.1853.v1

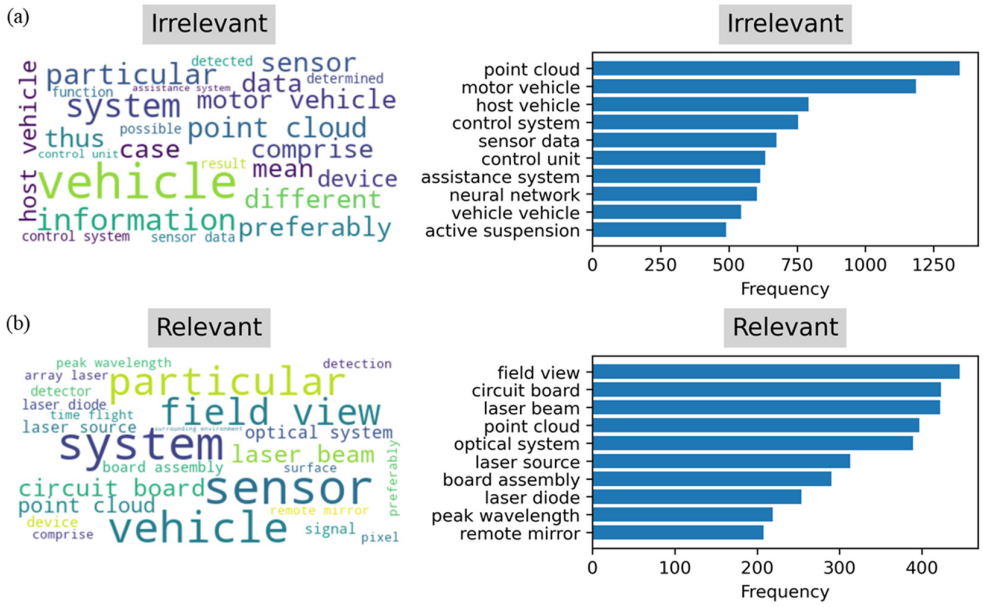

Figure 10. LiDAR top bigrams in a) irrelevant and b) relevant documents.

Error! Reference source not found. illustrates the term co-occurrence networks for irrelevant and relevant document sets within the corpus. **Error! Reference source not found.**a represents the irrelevant document set, where nodes like “vehicle,” “system,” and “sensor” dominate the network. The clusters reflect general topics, including automotive systems, sensors, and imaging technologies, which are indirectly related to LiDAR-specific designs. The loosely connected clusters, such as those involving “camera” and “map,” further indicate a broad and unfocused thematic scope.

Error! Reference source not found.b visualizes the relevant document set, showcasing a more specialized and interconnected network. Central nodes such as “LiDAR,” “laser,” and “system” dominate, with clusters focused on design topics specific to LiDAR, including “laser beam,” “point cloud,” “wavelength,” and “mirror.” These connections emphasize the core technologies and applications of LiDAR, particularly its optical and laser-based functionality for autonomous systems and environmental mapping.

 doi:10.20944/preprints202501.1853.v1

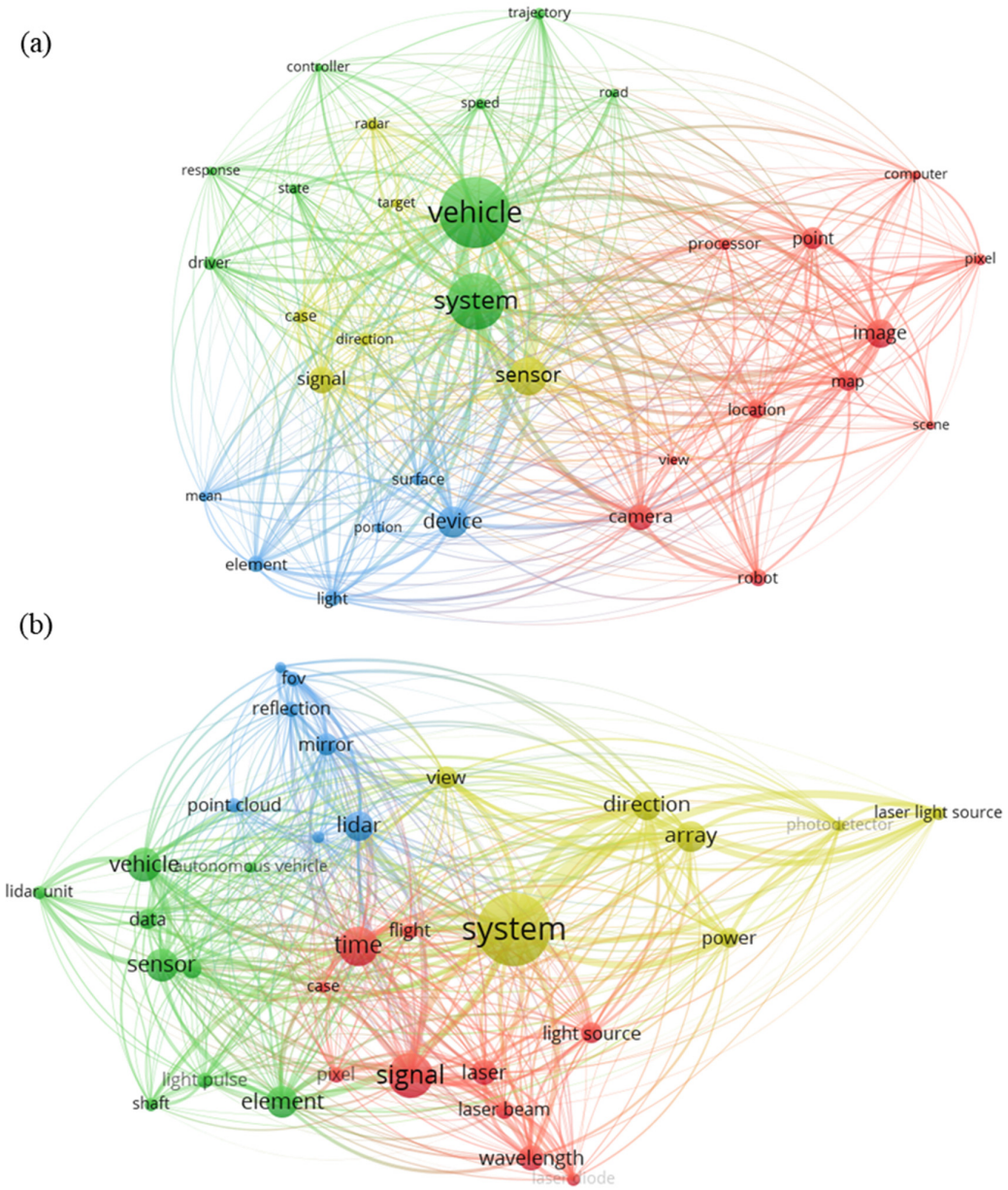

Figure 11. LiDAR term co-occurrence for a) irrelevant and b) relevant documents.

3.3. Unsupervised ML

The subsections that follow define the two sets of UML methods: two NLP topic modeling methods, and a manifold feature embedding method (t-SNE) with k-means clustering.

3.3.1. NLP Topic Modeling

Two NLP topic modeling methods, the types of UML widely used to uncover latent themes in text data, are latent Dirichlet allocation (LDA) and non-negative matrix factorization (NMF). LDA is a probabilistic generative model that characterizes each document as a mixture of topics, and each topic is a distribution over words in the corpus [18]. Intuitively, a word frequency distribution defines each of a predetermined number of topics, and a probability distribution of these topics characterizes the topical essence of each document. Mathematically, LDA maximizes the joint probability

 doi:10.20944/preprints202501.1853.v1

$$P(W, Z, \theta, \phi) = \prod_{d=1}^{D} P(\theta_d) \prod_{n=1}^{N_d} P(z_{d,n}|\theta_d) P(w_{d,n}|z_{d,n}, \phi) \quad (1)$$

where W represents the words in the corpus, Z represents the topic assignments, θ_d is the topic distribution for document d, ϕ is the word distribution for topics, $w_{d,n}$ is the nth word in document d, and $z_{d,n}$ is the topic assignment for that word. The algorithm used Gibbs sampling to learn the parameters θ and ϕ [19].

NMF, on the other hand, is a deterministic method based on matrix factorization [20]. Given a document-term matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, where m is the number of documents and n is the vocabulary size, NMF approximates $\boldsymbol{X}$ as the product of two non-negative matrices

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H} \quad (2)$$

where $\boldsymbol{W} \in \mathbb{R}^{m \times k}$ represents the topic weights for each document, $\boldsymbol{H} \in \mathbb{R}^{k \times n}$ represents the word weights for each topic, and k is the number of topics. The optimization objective is to minimize the Frobenius (F) norm

$$\min_{\boldsymbol{W},\boldsymbol{H}} \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2 \quad (3)$$

subject to the constraints that $\boldsymbol{W} \geq 0$ and $\boldsymbol{H} \geq 0$. This optimization utilized the method of coordinate descent [21].

The present research applied both LDA and NMF topic modeling methods to each of the five technology domains with the aim of separating relevant documents from irrelevant ones. The models automatically labeled the resulting topics as "topic 0" and "topic 1" to distinguish them. To evaluate the accuracy of these topic assignments, the workflow produced a confusion matrix using the labeled dataset, with relevant documents labeled as "1" and irrelevant documents labeled as "0." The confusion matrix served to assess the alignment between the predicted topic labels and the actual relevance labels. It summarized the counts of true positives (TP), which are relevant documents correctly assigned to the "relevant" topic, true negatives (TN), which are irrelevant documents correctly assigned to the "irrelevant" topic, false positives (FP), which are irrelevant documents incorrectly assigned to the "relevant" topic, and false negatives (FN), which are relevant documents incorrectly assigned to the "irrelevant" topic.

To refine the performance evaluation, the workflow calculated additional metrics such as precision, recall, and F1-scores, based on the confusion matrix. These metrics provide deeper insights into a model's ability to correctly classify relevant and irrelevant documents. Precision measures the proportion of documents predicted as relevant (positive) that are relevant, where

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (4)$$

High precision indicates low false positive rates, meaning the model is good at avoiding the misclassification of irrelevant documents as relevant. Recall, also known as sensitivity or true positive rate, measures the proportion of actual relevant documents that the model correctly identified, where

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (5)$$

High recall indicates that the model captures most of the relevant documents but does not incur a penalty for incorrectly classifying irrelevant ones.

The F1-score is the harmonic mean of precision and recall, providing a balanced metric that takes both FP and FN into account, where

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (6)$$

 doi:10.20944/preprints202501.1853.v1

The F1-score is particularly useful when there is an imbalance between the number of relevant and irrelevant documents. A high F1-score indicates that the model offers a good trade-off between correctly identifying relevant documents and avoiding false positives.

Since LDA and NMF do not inherently associate "topic 0" or "topic 1" with relevance or irrelevance, the analysis calculated the confusion matrix for both mappings of topics to relevance labels—that is "topic 0" as relevant and "topic 1" as irrelevant, and vice versa. The analysis selected the better result (higher classification accuracy) for evaluation. This approach ensured that judging the model's performance was based on the most appropriate alignment of topics with the labeled data, allowing for a fair and effective assessment of its ability to distinguish between relevant and irrelevant documents.

3.3.2. t_SNE and Clustering

t-SNE is a popular manifold learning method designed to embed high-dimensional data into a low-dimensional space, typically 2D, while preserving the local structure of the data [22]. Given a dataset with high-dimensional features, such as keyword frequency and first occurrence positions, t-SNE minimizes the divergence between two probability distributions: one that measures pairwise similarities in the high-dimensional space and another in the reduced space. The objective function of t-SNE minimizes the Kullback-Leibler divergence

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \, log \frac{p_{ij}}{q_{ij}} \tag{7}$$

where p_{ij} represents the joint probability of similarity between points i and j in the high-dimensional space, and q_{ij} represents the similarity in the low-dimensional space, modeled using Student's t-distribution. By projecting the feature set into two dimensions, t-SNE enables intuitive visualization and downstream clustering.

After reducing the features to two dimensions with t-SNE, the workflow applied k-means clustering, a type of UML model, to group the documents into two clusters, C0 and C1. The algorithm aimed to minimize intra-cluster variance by iteratively updating cluster centroids and assigning data points based on Euclidean distance. The analysis evaluated the performance of this clustering method using a confusion index (*CI*), defined as

$$CI = \frac{\min(n_{C0}, n_{C1})}{\max(n_{C0}, n_{C1})} \tag{8}$$

where n_{C0} and n_{C1} are the number of relevant documents in C0 and C1, respectively. Hence, a lower *CI* indicates less confusion, meaning one cluster mostly contains relevant documents while the other mostly contains irrelevant documents. A perfect bifurcation of documents would yield $CI = 0$, indicating complete separation of relevant and irrelevant documents. This method provides an interpretable way to assess the relevance filtering performance of feature-based clustering, with the visualization from t-SNE aiding in identifying separable clusters and their alignment with the labeled relevance.

3.4. Supervised ML

3.4.1. Statistical Models

The workflow evaluated several SML models for classifying documents as relevant or irrelevant. Each method offers unique strengths, hence, comparing their performance determined the most effective approach for this task. These SML algorithms, applied to the labeled dataset, provide diverse approaches to classification, allowing a comprehensive comparison of their effectiveness and computational efficiency. The following are brief descriptions for each of the 10 algorithms evaluated:

Gradient Boosting (GB): An ensemble method that builds decision trees sequentially, with each tree reducing the errors of its predecessor by minimizing a specified loss function. For predictions, the ensemble output is

 doi:10.20944/preprints202501.1853.v1

$$\hat{y} = \sum_{t=1}^{T} \eta f_t(x) \tag{9}$$

where x is the set of features, $f_t(x)$ is the tth decision tree, η is the learning rate, and T is the total number of trees.

Extreme Gradient Boosting (XGB): A more efficient implementation of GB that incorporates regularization to prevent overfitting and parallel processing to handle large datasets effectively.

Random Forest (RF): An ensemble method that builds multiple decision trees using random subsets of features and data. The final prediction is the majority vote for classification

$$\hat{y} = \text{mode}\{f_t(x)\}_{t=1}^{T} \tag{10}$$

where $f_t(x)$ is the prediction from the tth tree.

Naïve Bayes (NB): A probabilistic classifier based on Bayes' theorem, assuming feature independence. The posterior probability for class C_k is

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} \tag{11}$$

where $P(x|C_k)$ is the likelihood, $P(C_k)$ is the prior, and $P(x)$ is the evidence.

AdaBoost (ADB): An adaptive boosting method that combines weak classifiers sequentially, adjusting weights for misclassified examples to focus on difficult cases.

k-Nearest Neighbor (kNN): A non-parametric algorithm that assigns a class to a data point based on the majority class of its k nearest neighbors in feature space.

Decision Tree (DT): A tree structure where nodes split data based on feature thresholds to maximize information gain or minimize Gini impurity. Predictions are based on the majority class in a leaf node.

Support Vector Machine (SVM): A linear classifier that finds the hyperplane maximizing the margin between two classes. For linear SVM, the decision boundary satisfies

$$w^T x + b = 0 \tag{12}$$

where w is the weight vector and b is the bias term.

Logistic Regression (LR): A probabilistic model that estimates the likelihood of a class using the sigmoid function

$$P(y = 1|x) = \frac{1}{1 + e^{-(w^T x + b)}} \tag{13}$$

where w is the weight vector and b is the bias term.

Stochastic Gradient Descent (SGD): A linear SVM implemented with the hinge loss function

$$\text{Loss} = \max\big(0, 1 - y(w^T x + b)\big) \tag{14}$$

where y is the true label and x is the feature vector. SGD iteratively updates the weights to minimize this loss.

3.4.2. Deep Learning

D-ANN represents a class of SML models inspired by the structure of the human brain. Unlike classical SML models, which often rely on manually designed features, D-ANNs automatically learn hierarchical feature representations from raw data, making them particularly powerful for complex, high-dimensional datasets. D-ANNs consist of multiple layers of interconnected nodes (neurons), including an input layer, hidden layers, and an output layer. Each neuron computes a weighted sum of its inputs, applies an activation function (e.g., ReLU or sigmoid), and passes the result to the next layer [23]. The output of the network is determined by

 doi:10.20944/preprints202501.1853.v1

$$y = f\left(W_L \cdot \sigma\left(W_{L-1} \cdot \sigma\left(\cdots \cdot \sigma(W_1 \cdot x + b_1) + b_{L-1}\right)\right) + b_L\right) \quad (15)$$

where x is the input, W_l and b_l are the weights and biases for layer l, σ is the activation function, and f is the output layer activation. Heuristically designed, the least complex model that provided diminishing returns was two fully connected inner layers of 100 nodes each, with the SoftMax function utilized for activation. **Error! Reference source not found.** contrasts the characteristics of D-ANN with the statistical SML models.

Table 3. Characteristics of deep-learning and statistical ML models.

Category	D-ANN SML	Statistical SML
Data Features	Automatically learns features directly from high-dimensional and large-scale data, enabling them to capture nonlinear relationships in the data.	Rely on engineered features and have difficulty capturing nonlinear relationships in high-dimensional data.
Performance	Achieves state-of-the-art results, particularly in image, speech, and text processing.	Less prone to overfitting from pattern memorization but have more difficulty learning representations from noisy data.
Computational Cost	Requires significant computational resources such as special purpose processors to support their longer training time.	Computationally efficient and faster to train.
Scalability	Excel in scenarios with abundant data, leveraging their capacity to model complex relationships.	Typically perform well with smaller and cleaner datasets.
Interpretability	Often described as “black boxes” because their predictions are more difficult to interpret.	Simpler models like decision trees and linear models provide insights into their representation or explanation of the data.

This study applied D-ANN to the labeled dataset and compared its performance with nine classical SML methods. The analysis evaluated their classification accuracy, computational cost, and robustness in filtering relevant documents.

To evaluate the performance of the SML models, the workflow used k-fold cross-validation to produce three key evaluation metrics: area under the curve (AUC), classification accuracy (CA), and F1-score. *K*-fold cross-validation is a robust validation technique that splits the dataset into *k* equally sized “folds” or subsets. The procedure trained the models on *k*-1 folds and tested them on the remaining fold. The procedure repeated the process *k* times, with each fold serving as the test set once. The final performance metrics were their average across all *k* iterations, ensuring a reliable and unbiased evaluation by mitigating the effects of data partitioning. Iteratively adjusting the number of folds revealed that three folds provided the best results across all models.

AUC measures the area under the receiver operating characteristic (ROC) curve, which plots the TP rate (sensitivity) against the FP rate (1-specificity) at various threshold settings [24]. AUC values range from 0 to 1, with higher values indicating better discriminatory ability of the model to distinguish between relevant and irrelevant documents. CA represents the proportion of correctly classified documents, regardless of their relevance. Hence, CA provides an overall measure of how well the model performs but does not account for class imbalances. All models had some default hyperparameters that were not part of the training process. The workflow employed an iterative process to adjust hyperparameters in small increments on either side of the default values to maximize their average AUC performance.

 doi:10.20944/preprints202501.1853.v1

4. Results

The following subsections discuss the results of the UML (NLP topic modeling, t-SNE clustering) and SML methods evaluated for document relevance classification.

4.1. NLP Topic Modeling

Error! Reference source not found. summarizes the performance metrics of the topic modeling. The "Best" parameter indicates that LDA did better than NMF for all categories, based on the F1-score. The "-T" or "-F" after the model abbreviation indicates "True" or "False," respectively, if swapping the default topic assignments resulted in better model performance.

Table 4. Topic modeling performance for each technology domain.

Parameter	SSB	EVC	CV	eVTOL	LiDAR	Mean
Best	LDA-T	LDA-T	LDA-T	LDA-T	LDA-F	
TN	27	412	205	542	402	
TP	179	218	144	125	170	
FN	68	172	76	33	21	
FP	7	382	175	713	143	
Precision	0.96	0.36	0.45	0.15	0.54	0.49
Recall	0.72	0.56	0.65	0.79	0.89	0.72
F1	0.83	0.44	0.53	0.25	0.67	0.55

The LDA model with the default topic-to-relevance mapping ("-T") yielded the highest F1-score for four of the five categories (SSB, EVC, CV, and eVTOL), while a flipped topic assignment ("-F") resulted in better performance for the LiDAR category. Key observations are that SSB achieved the highest precision (0.96), indicating very few false positives. Conversely, eVTOL had the lowest precision (0.15), reflecting significant misclassification of irrelevant documents as relevant. LiDAR achieved the highest recall (0.89), showing the model's ability to capture most of the relevant documents. SSB had a comparatively lower recall (0.72), indicating that the model missed some relevant documents. SSB had the highest F1-score (0.83), demonstrating strong overall performance. eVTOL had the lowest F1-score (0.25), primarily due to poor precision.

The average precision across all categories was 0.49, indicating that topic modeling yielded moderate performance in identifying relevant documents. The average recall was higher at 0.72, suggesting the model was better at identifying relevant documents. The average F1-score was 0.55, reflecting a moderate trade-off between precision and recall. LDA consistently outperformed NMF in identifying relevant documents, with default topic-to-relevance mappings often providing the best results. The models provided the best overall performance on the SSB corpus, with high precision, recall, and F1-score, indicating well-separated topics by this method. The models performed the worst on the eVTOL corpus due to high false positives, indicating challenges in separating relevant from irrelevant topics in this technology domain. The model showed high recall on the LiDAR corpus, suggesting good coverage of relevant documents but at the cost of precision. These results highlight the strengths and weaknesses of topic modeling for relevance filtering across various technology domains and suggest that LDA is often better suited for this task than NMF, albeit with variability depending on the dataset characteristics.

4.2. t-SNE Clustering

Error! Reference source not found. illustrates the results of clustering documents based on the t-SNE reduction of features derived from keyword frequency and first position of occurrence in each document.

 doi:10.20944/preprints202501.1853.v1

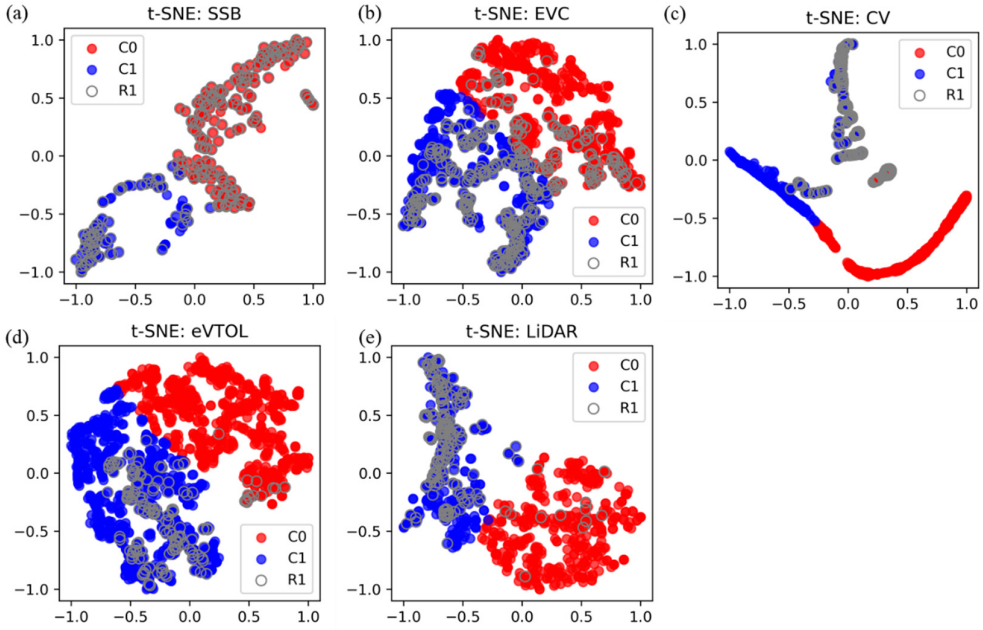

Figure 12. Visualization of the confusion index from the unsupervised ML for each technology domain.

Each subfigure (a–e) corresponds to one of the five technology domains. The t-SNE method reduced the features to two dimensions to facilitate visualization of the clustering performed by k-means. Each point in the plot represents a single document in the normalized two-dimensional t-SNE space. As indicated in the legend, red and blue dots correspond to the two clusters (C0, C1) identified by the *k*-means algorithm. Gray circles mark documents labeled as relevant (R1) in the dataset, providing a visual representation of the confusion between relevant and irrelevant classifications. That is, significant overlap of gray circles across both clusters reflects higher confusion, whereas concentration in one cluster indicates better performance.

Error! Reference source not found. summarizes the quantitative results of the clustering for each technology domain.

Table 5. Classification based on keyword frequency and first position features.

Parameter	SSB	EVC	CV	eVTOL	LiDAR
C0	184	608	276	645	408
C1	97	576	324	768	328
R1C0	175	131	38	16	15
R1C1	72	259	182	142	176
CI	0.41	0.51	0.21	0.11	0.09

Key parameters include:

- C0 and C1: Total number of documents assigned to Cluster 0 (C0) and Cluster 1 (C1).
- R1C0 and R1C1: Number of relevant documents in Cluster 0 and Cluster 1, respectively.
- Confusion Index (CI): Metric defined above to measure clustering confusion.

Lower CI values indicate better separation of relevant documents into a single cluster, while higher values signify greater confusion. For the SSB case, relevant documents spread evenly across the clusters, resulting in a high CI of 0.41. This indicates significant confusion in distinguishing relevant documents from irrelevant ones based on the chosen features. For the EVC case, relevant documents are similarly distributed, resulting in a high CI of 0.51. For the CV case, documents concentrate in Cluster 1, leading to a low CI of 0.21. Hence, the separation of relevant documents improves significantly compared with SSB and EVC. For the eVTOL case, the distribution is uneven,

 doi:10.20944/preprints202501.1853.v1

leading to a low CI of 0.11, and showing good clustering performance. For the LiDAR case, relevant documents concentrate in Cluster 1, yielding the lowest CI of 0.09.

Overall, the t-SNE and k-means approach performs better for the CV, eVTOL, and LiDAR technology domains, as indicated by lower CI values, suggesting effective separation of relevant documents. In contrast, SSB and EVC exhibit higher CI values, indicating significant confusion between clusters. The features (keyword frequency and first position of occurrence) seem more effective for separating relevant documents in domains with distinct terminologies, such as eVTOL and LiDAR. Domains with overlapping or broad terminologies, such as SSB and EVC, show greater clustering challenges. These results highlight the variability of clustering effectiveness based on the extracted keyword features across different technology domains. The outcome emphasizes the importance of feature selection and domain-specific considerations when using UML techniques for relevance filtering.

4.3. Supervised ML

Error! Reference source not found. presents the results of SML approaches for classifying topic relevance using features based on keyword frequency and first position of occurrence. The best-performing model for each technology domain is based on the highest AUC score. The parameter "No. AUC90" indicates the number of models achieving an AUC of at least 0.90, reflecting the robustness of performance across models. The parameter "C(ANN/Best)" represents the computational time ratio between D-ANN and the best-performing classical SML model, or the next best model if D-ANN was the best.

Table 6. Classification based on supervised ML.

Parameter	SSB	EVC	CV	eVTOL	LiDAR	Mean
Best Model	NB	GB	D-ANN	GB	XGB	
AUC	0.806	0.998	0.994	0.987	0.948	0.95
CA	0.890	0.981	0.983	0.962	0.901	0.94
F1	0.876	0.981	0.983	0.963	0.899	0.94
No. AUC90	0	8	10	9	7	6.8
C(ANN/Best)	53.5	13.4	18.1	15.1	23.7	24.8

The results highlight several key findings:

- Classical SML models, such as NB, GB, and XGB, yielded comparable performance to the D-ANN in terms of AUC, CA, and F1 scores. For instance, GB was the best model for EVC and eVTOL, achieving an average AUC of 0.95 across all technology domains.
- D-ANN achieved the highest performance for CV (AUC = 0.994) but required significantly more computational resources, with a C(ANN/Best) ratio averaging 24.8 across the technology domains. In some cases, such as SSB, the D-ANN required 53.5 times more computational time than the best-performing model (NB), despite similar classification performance.
- The SML approach outperformed UML methods (topic modeling and t-SNE clustering) in all metrics, demonstrating superior ability to classify topic relevance in patent documents covering diverse technical domains.
- The number of models achieving AUC ≥ 0.90 ("No. AUC90") was consistently high, particularly for CV (10) and eVTOL (9), indicating that SML methods are robust across those feature sets and technical domains.

Overall, these results demonstrate that while D-ANN models can deliver exceptional performance, statistical SML models can achieve similar performance with significantly lower computational demands. This makes statistical SML models more practical for applications with limited computational resources. Furthermore, the marked improvement of SML methods over UML approaches highlights their effectiveness in relevance classification tasks across a variety of technical domains.

 doi:10.20944/preprints202501.1853.v1

5. Discussions

This study demonstrates the potential of ML methods to optimize document relevance filtering, particularly in multidisciplinary patent datasets. The results reveal critical insights into the strengths and weaknesses of different methodologies applied to the five technology domains. By systematically evaluating different methods of UML and SML, the study highlights their relative effectiveness in handling large and complex corpora with diverse terminologies.

The key insight from the results is that SML methods are superior in this application. SML models consistently outperformed UML methods in classifying document relevance. Metrics such as AUC, classification accuracy, and F1-score highlight the robustness of models like GB, XGB, and D-ANNs. The ability of SML models to leverage labeled data allowed for precise discrimination between relevant and irrelevant documents, particularly evident in domains with distinct terminologies like eVTOL and LiDAR. Topic modeling techniques (LDA and NMF) demonstrated moderate success, with SSB achieving the highest F1-score among the technology domains analyzed. UML experienced significant challenges in aligning topic labels with document relevance. Clustering using t-SNE and k-means provided mixed results. Lower confusion indices in domains like eVTOL and LiDAR reflect effective clustering, while higher confusion in the domains of SSB and EVC suggests that just clustering features such as keyword frequency and positional occurrence may not fully capture domain-specific nuances.

There were a few domain-specific observations. The SSB domain achieved the highest precision in UML methods, indicating the effectiveness of LDA in separating topics. However, SML models provided a more balanced trade-off between precision and recall. The LiDAR domain showed exceptional recall, suggesting high coverage of relevant documents but with the trade-off of increased false positives in UML methods. CV and eVTOL domains benefited significantly from SML approaches, where advanced models like D-ANN captured the complexity of interrelated features better than UML methods. However, D-ANN models required significantly higher computational resources compared with the statistical SML methods.

The limitations of this study stem primarily from its scope and methodology. The literature search focusing on methods of patent document relevance filtering returned very few recent studies. The UML methods relied on basic features such as keyword frequency and positional occurrence. While effective for certain domains, these features struggled to encapsulate domain-specific complexities, particularly for SSB and EVC. Although those features worked well for the SML models, they require labeled datasets, which may not be readily available for emerging domains or new datasets. Future work will explore richer feature sets, including semantic embeddings and contextual word representations. The focus on specific transportation-related domains limits the broader applicability of the findings. While the results are encouraging, future studies will validate these methods across other distinct fields, such as healthcare and renewable energy.

6. Conclusions

This study addressed the critical challenge of filtering relevant documents from large patent datasets in multidisciplinary technology domains. By systematically evaluating both supervised and unsupervised machine learning methods, the research highlights the strengths and weaknesses of various approaches. Supervised methods, particularly gradient boosting, extreme gradient boosting, and deep artificial neural networks, emerged as the most effective for accurately classifying document relevance, offering high precision, recall, and F1-scores across diverse technology domains. Unsupervised methods, while valuable for exploratory analysis of unlabeled data, exhibited variability in performance depending on domain-specific characteristics and feature engineering.

The findings highlight the importance of domain-specific considerations and robust feature selection in relevance filtering tasks. They also demonstrate the trade-offs between computational efficiency and model accuracy, with statistical supervised models providing a practical balance for

 doi:10.20944/preprints202501.1853.v1

resource-constrained applications. The study advances knowledge by offering a benchmark for document filtering based on machine learning methods, enabling researchers and practitioners to optimize their data curation processes.

Future work will focus on incorporating advanced feature representations and expanding datasets to validate the methodologies across broader applications. By addressing these areas, research can further advance the scalability and generalizability of document filtering driven by machine learning methods, contributing to more efficient research, planning, and innovation across various fields.

Funding: This research received funding from the United States Department of Transportation, Center for Transformative Infrastructure Preservation and Sustainability (CTIPS), Funding Number 69A3552348308.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: This article includes the data presented in the study.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. J. Lee, S. Park and J. Lee, "A Fast and Scalable Algorithm for Prior Art Search," *IEEE Access,* vol. 10, pp. 7396-7407, 2022.
2. A. Ali, A. Tufail, L. C. D. Silva and P. E. Abas, "Innovating patent retrieval: a comprehensive review of techniques, trends, and challenges in prior art searches," *Applied System Innovation,* vol. 7, no. 5, p. 91, 2024.
3. W. Shalaby and W. Zadrozny, "Patent Retrieval: A Literature Review," *Knowledge and Information Systems,* vol. 61, pp. 631-660, 2019.
4. J. Risch and R. Krestel, "Domain-specific Word Embeddings for Patent Classification," *Data Technologies and Applications,* vol. 53, no. 1, pp. 108-122, 2019.
5. A. J. Trappey, C. V. Trappey, J.-L. Wu and J. W. Wang, "Intelligent Compilation of Patent Summaries Using Machine Learning and Natural Language Processing Techniques," *Advanced Engineering Informatics,* vol. 43, p. 101027, 2020.
6. J. Just, "Natural Language Processing For Innovation Search–Reviewing An Emerging Non-Human Innovation Intermediary," *Technovation,* vol. 129, p. 102883, 2024.
7. G. Kumaravel and S. Sankaranarayanan, "PQPS: Prior-Art Query-Based Patent Summarizer Using RBM and Bi-LSTM," *Mobile Information Systems,* p. 2497770, 2021.
8. S. Casola and A. Lavelli, "Summarization, Simplification, and Generation: The Case of Patents," *Expert Systems with Applications,* vol. 205, p. 117627, 2022.
9. G. Yue, J. Liu, Y. Hou and Q. Zhang, "A Novel Patent Knowledge Extraction Method for Innovative Design," *IEEE Access,* vol. 11, pp. 2182-2198, 2022.
10. C. Li, W. Li, Y. Hong and H. Xiang, "A Patent Retrieval Method and System Based on Double Classification," *Information Sciences,* vol. 672, p. 120659, 2024.
11. H. Wu, G. Shen, X. Lin, M. Li, B. Zhang and C. Z. Li, "Screening patents of ICT in construction using deep learning and NLP techniques," *Engineering, Construction and Architectural Management,* vol. 27, no. 8, pp. 1891-1912, 2020.
12. R. Krestel, R. Chikkamath, C. Hewel and J. Risch, "A Survey on Deep Learning for Patent Analysis," *World Patent Information,* vol. 65, p. 102035, 2021.
13. E. Sulis, L. Humphreys, F. Vernero, I. A. Amantea, D. Audrito and L. D. Caro, "Exploiting Co-Occurrence Networks for Classification of Implicit Inter-Relationships in Legal Texts," *Information Systems,* vol. 106, p. 101821, 2022.
14. J. T. Pintas, L. A. Fernandes and A. C. B. Garcia, "Feature Selection Methods for Text Classification: A Systematic Literature Review," *Artificial Intelligence Review,* vol. 54, no. 8, pp. 6149-6200, 2021.

 doi:10.20944/preprints202501.1853.v1

15. R. Krestel, R. Chikkamath, C. Hewel and J. Risch, "A Survey on Deep Learning for Patent Analysis," *World Patent Information,* vol. 65, p. 102035, 2021.
16. PTMT, "Patent Technology Monitoring Team (PTMT)," 2020. [Online]. Available: https://www.uspto.gov/web/offices/ac/ido/oeip/taf/reports.htm. [Accessed 9 January 2025].
17. USPTO, "U.S. Patent and Trademark Office (USPTO), Patent Technology Monitoring Team," 30 May 2021. [Online]. Available: https://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm. [Accessed 9 January 2025].
18. J. Zimmermann, L. E. Champagne, J. M. Dickens and B. T. Hazen, "Approaches to Improve Preprocessing for Latent Dirichlet Allocation Topic Modeling," *Decision Support Systems,* vol. 185, 2024.
19. A. Agrawal, W. Fu and T. Menzies, "What is wrong with topic modeling? And how to fix it using search-based software engineering.," *Information and Software Technology,* vol. 98, pp. 74-88, 2018.
20. G. R. Naik, Ed., Non-negative Matrix Factorization Techniques: Advances in Theory and Applications, Heidelberg: Springer, 2016.
21. S. J. Wright, "Coordinate Descent Algorithms," *Mathematical Programming,* vol. 151, pp. 3-34, 2015.
22. F. Anowar, S. Sadaoui and B. Selim, "Conceptual and Empirical Comparison of Dimensionality Reduction Algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, T-SNE)," *Computer Science Review,* vol. 40, p. 100378, 2021.
23. C. C. Aggarwal, Data Mining, New York, New York: Springer International Publishing, 2015, p. 734.
24. A. Burkov, The Hundred-Page Machine Learning Book, Quebec City: Andriy Burkov, 2019.