applied sciences

MDPI

Article

Alignment of Unsupervised Machine Learning with Human Understanding: A Case Study of Connected Vehicle Patents

Raj Bridgelall

Transportation, Logistics, & Finance, College of Business, North Dakota State University, P.O. Box 6050, Fargo, ND 58108-6050, USA; raj@bridgelall.com or raj.bridgelall@ndsu.edu

Featured Application: Technology trend analysis of connected vehicles.

Abstract: As official public records of inventions, patents provide an understanding of technological trends across the competitive landscape of various industries. However, traditional manual analysis methods have become increasingly inadequate due to the rapid expansion of patent information and its unstructured nature. This paper contributes an original approach to enhance the understanding of patent data, with connected vehicle (CV) patents serving as the case study. Using free, open-source natural language processing (NLP) libraries, the author introduces a novel metric to quantify the alignment of classifications by a subject matter expert (SME) and using machine learning (ML) methods. The metric is a composite index that includes a purity factor, evaluating the average ML conformity across SME classifications, and a dispersion factor, assessing the distribution of ML assigned topics across these classifications. This dual-factor approach, labeled the H-index, quantifies the alignment of ML models with SME understanding in the range of zero to unity. The workflow utilizes an exhaustive combination of state-of-the-art tokenizers, normalizers, vectorizers, and topic modelers to identify the best NLP pipeline for ML model optimization. The study offers manifold visualizations to provide an intuitive understanding of the areas where ML models align or diverge from SME classifications. The H-indices reveal that although ML models demonstrate considerable promise in patent analysis, the need for further advancements remain, especially in the domain of patent analysis.

Keywords: natural language processing; NLP model evaluation; machine learning; patent classification; semantic analysis; human-centered interpretability; technological trend analysis

check for updates

Citation: Bridgelall, R. Alignment of Unsupervised Machine Learning with Human Understanding: A Case Study of Connected Vehicle Patents. *Appl. Sci.* **2024**, *14*, 474. https://doi.org/10.3390/app14020474

Academic Editor: Andrea Prati

Received: 8 December 2023
Revised: 2 January 2024
Accepted: 3 January 2024
Published: 5 January 2024

1. Introduction

Patents are a valuable source of information for understanding technological advancements, identifying emerging trends, and assessing the competitive landscape of various industries. However, because of the vast and growing volume of patent data and their unstructured nature, they are challenging to analyze manually [1]. Current approaches in patent analysis have primarily focused on manual interpretations, leading to challenges in scalability and consistency. This paper addresses that gap by evaluating the use of natural language processing (NLP) and machine learning (ML) techniques to extract insights from patent documents, particularly in the context of connected vehicle (CV) patents.

Analysts have used NLP and ML to automate tasks such as patent classification, topic modeling, technology identification, and patent recommendation systems [2]. Even so, verifying the alignment of ML methods with human understanding of patent documents remains challenging and is an unresolved issue [3]. Hence, establishing a performance benchmark of state-of-the-art, free, open-source models will set the stage to evaluate the advancement of machines that can understand and classify patents. Such advancements could include large language models (LLMs), but accessing their application programming interface (API) is not currently free or widely available to everyone, and the models are still immature.

The goal of this research is to develop a comprehensive metric for quantifying alignment between subject matter expert (SME) and ML classification of patent topics. The author selected the field of CVs for the case study based on his more than 30 years of domain knowledge and industry experience inventing relevant products. With dozens of related patent awards in the field, the author amassed decades of experience reviewing and analyzing the nuances of patent documents in the CV and intelligent transportation systems domain. Hence, the author is qualified to serve as the SME in this study.

The contribution of this paper is an original metric to quantify the alignment between SME classification and ML topic assignments. The proposed composite index comprises a purity factor that measures the average purity across SME classifications, and a dispersion factor that measures the spread of assigned topics across the SME classifications. The analysis of various mature methods provides valuable insights into the effectiveness of different combinations of tokenizers, normalizers, vectorizers, and topic modelers in the NLP pipeline. The embedded manifold visualizations effectively demonstrate the areas where the topic model aligns and diverges from the SME classifications, offering a clear view of the model's performance and limitations. This paper also bridges a gap in the literature between NLP efficiency and human-centered interpretability in topic modeling.

The organization of the rest of this paper is as follows: Section 2 reviews the literature on CV technology development, the utility of NLP in patent analysis, and evaluations of NLP alignment with human judgment. Section 3 describes the data preparation, the NLP pipeline, and optimization of the selected topic model. Section 4 presents the results, including visualizations of the confusion between ML models and human interpretations. Section 5 discusses implications, further insights, and limitations of the work. Section 6 concludes the research and suggests future work.

2. Literature Review

The three subsections of the literature review focus on connected vehicle trends, NLP in patent analysis, and quantifying the alignment of ML methods with human judgment.

2.1. Connected Vehicles Trends

A review of the latest trends in CV technology development sets the stage for understanding the topic classifications. The transportation industry expects CV technology to revolutionize driving safety, efficiency, and convenience [4]. The recent literature on CVs focused on advancing security, traffic management, and cooperative control systems. Nkenyereye et al. (2023) examined the integration of 5G networks in vehicular cloud computing, highlighting the potential of vehicle clusters to share resources and data in a mobile cloud environment [5].

The security of CVs is another area of major concern. Shichun et al. (2023) reviewed cybersecurity techniques focusing on CVs, covering threat analysis and intrusion detection [6]. Rathore et al. (2023) focused on the cybersecurity challenges of in-vehicle communication [7], while Ju et al. (2022) examined attack detection and resilience from a vehicle control perspective [8]. Hildebrand et al. (2023) explored the use of blockchain technology to enhance security in vehicle networking [9], and Khan et al. (2023) proposed a blockchain-based secure communication system for CVs [10].

Alanazi (2023) conducted a systematic literature review of how autonomous vehicles manage traffic at junctions, exploring various methodologies that include ML [11]. Shi et al. (2023) developed a real-time control algorithm for CVs to optimize fuel use in signalized corridors [12]. Gholamhosseinian and Seitz (2022) surveyed cooperative intersection management strategies for CVs [13]. Xu and Tian (2023) proposed a method to improve arterial signal coordination using CV data [14], and Zhu et al. (2022) reviewed merging control strategies at freeway on-ramps [14]. Wang et al. (2022) discussed the development of cooperative driving systems for CVs [15].

Cui et al. (2022) reviewed cooperative perception technologies that combine local and edge sensing data for improved situational awareness [16]. Khanal et al. (2023) utilized CV data to develop crash prediction models. Gao et al. (2023) provided insights into predictive cruise control under cloud control systems, emphasizing the role of predictive algorithms in traffic efficiency and safety [17]. Islam and Abdel-Aty (2023) focused on using CV data for traffic conflict prediction [18]. Schwarz et al. (2022) examined the role of digital twin simulations in various traffic management applications [19].

The above literature review revealed that trending topics in CV development include vehicular security, traffic management, cooperative control and perception systems, and overall driving safety. These trends suggest that while the industry has made considerable progress in the development of technologies for CVs, challenges in security, real-time control, and effective traffic management remain, necessitating further research and innovation in these areas.

2.2. NLP in Patent Analysis

This section of the literature review synthesizes NLP-related contributions in patent analysis from the recent literature and contrasts them with the current paper's unique contributions. Krestel et al. (2021) surveyed the use of deep learning for patent analysis, emphasizing the significant role of these techniques in automating tasks previously solvable only by domain experts [2]. Casola and Lavelli (2022) surveyed NLP approaches for summarizing, simplifying, and generating patent text for non-experts to improve the accessibility and understanding of patent information for a wider audience [1]. Their work highlighted the peculiar challenges patents pose to current NLP systems. This finding suggested that the performance of NLP and ML methods is sensitive to the type of topic domain analyzed.

Trappey et al. (2020) focused on NLP-based patent information retrieval and intelligent compilation of patent summaries [20]. Joshi et al. (2022) proposed an intelligent keyword extraction technique for patent classification by training a transformer model and comparing its performance with K-means clustering and topic modeling using the latent Dirichlet allocation (LDA) method [21]. These works highlight the challenges in effectively extracting and summarizing knowledge from patents and the opportunities that NLP techniques offer in this context.

De Clercq et al. (2019) proposed a multi-label classification approach to classify electric vehicle patents by combining LDA with ML algorithms to explore the relationships between patents and cooperative patent classification classes [22]. Their method provided a user-friendly way to analyze and visualize patent data. Hyun et al. (2020) and Wu et al. (2020) explored semantic analysis of patent data, emphasizing the role of NLP in identifying technical trends [23] and screening patents in specific domains such as communication technologies in construction [24]. These studies highlight the evolution of NLP and ML tools and the need for continuous innovation in this field.

Arts et al. (2021) demonstrated the potential of topic modeling in enhancing the understanding of patent documents and providing insights into the evolution of technologies [25]. Puccetti et al. (2023) proposed a named entity recognition (NER) method to identify technology-related entities from patent texts [26]. Their approach utilized a combination of rule-based and ML techniques to identify entities such as products, processes, organizations, and locations. Rezende et al. (2022) combined NLP techniques with algorithms such as latent semantic analysis (LSA), word2vec, and word mover's distance (WMD) to analyze patent similarity and technology trends [27].

2.3. NLP Alignment Evaluation

This subsection focuses on works that surveyed or compared the performance of NLP and ML methods. Kherwa and Bansal (2019) presented a comprehensive survey of topic modeling, highlighting challenges in their quantitative evaluation [28]. Abdelrazek et al. (2023) provided a recent survey categorizing topic modeling techniques into algebraic, fuzzy, probabilistic, and neural categories [29]. They reviewed the diversity and evolution of topic models and found that the research trends are moving toward developing and tuning neural topic models such as LLMs. Meaney et al. (2023) focused on methods for assessing the quality of topic models. They explored metrics such as reconstruction error, topic coherence, and stability analysis, emphasizing that different metrics capture various aspects of model fit [30]. Their findings suggested that a combination of indices, coupled with human validation, is essential for assessing the performance of topic models, a perspective that highlights the complexities of evaluating model quality. Harrando et al. (2021) empirically evaluated the challenges in systematically comparing topic modeling algorithms, revealing shortcomings in common practices and highlighting the need for a standardized approach in model benchmarking [31]. Vayansky and Kumar (2020) reviewed various topic modeling methods, focusing on their ability to handle complex data relationships, such as correlations between topics and topic changes over time. Their work encouraged diversity in the choice of topic modeling methods, particularly for complex datasets [32]. Borghesani et al. (2023) compared human and artificial semantic representations in topic modeling [3]. They demonstrated that NLP embeddings still fall short of human-like semantic representations. Rüdiger et al. (2022) compared non-application-specific topic modeling algorithms and assessed their performance against known clustering [33]. They found that metrics used so far provided a mixed picture that made it difficult to verify the accuracy of topic modeling outputs, concluding that topic model evaluation remains an unresolved issue [33]. Hoyle et al. (2021) questioned the validity of automated topic model evaluation, suggesting a disparity between automated coherence metrics and human judgment [34].

The above studies demonstrate the wide range of applications for NLP and ML in patent analysis, highlighting the complexities in evaluating algorithmic interpretations in topic modeling and the importance of human validation. Relative to these studies, the present study stands out by uniquely combining topic modeling techniques with SME expertise to quantify alignment in patent topic classification, particularly in the context of CV patents. The proposed topic alignment index, combined with an optimized NLP pipeline, fills a crucial gap in the literature by offering a more rigorous and complete approach in establishing a performance benchmark to contrast with future advancements. As NLP and ML techniques continue to advance, their role in patent analysis is likely to expand, providing researchers, practitioners, and policymakers with even more powerful tools for understanding, managing, and forecasting technological innovation. Hence, this study not only contributes to the existing body of knowledge, but also opens new avenues for future research in patent analysis.

3. Methodology

The methodology aims to quantify the alignment between patent topic classification by an SME and that using ML methods that are part of an NLP pipeline. The first subsection below describes the data preparation and the results of the SME classification. Four procedures make up the NLP pipeline, and several types of mature algorithms are available for each procedure. The second subsection describes the NLP pipeline in terms of the combinations of algorithms evaluated. As indicated in the literature review above, the existing literature does not define an objective method to quantify classification alignment. Hence, the third subsection describes the objective method developed in this work. The fourth subsection describes the method of identifying the best NLP pipeline and optimizing the identified topic model in terms of finding the optimum number of topics to set for the hyperparameter.

3.1. Data Preparation

This section details the data source selection, patent record extraction and selection, SME classification of distinct topic areas, and visualizations of the topic distributions and their dominant keywords. According to data from the World Intellectual Property

Organization (WIPO), the United States and China were the world's top patent holders in CV development [35]. Therefore, this research selected the United States Patent and Trademark Office (USPTO) as a representative baseline, focusing on the English language to classify patent topics in connected vehicle development [36]. A bigram search of the USPTO patent summary from 2018 to 2022 extracted 622 CV-related records from a total of 1,637,725. Selecting the last five full years (rather than partial data available for 2023) provided a more accurate representation of annual trends. The author, serving as the SME, subsequently identified 220 of these patents that were directly relevant to CV technology development.

The SME classification resulted in the 11 distinct topic areas summarized in Table 1. The table lists each topic and describes the general types of problems addressed within that category. Figure 1 shows a distribution of these topic areas by year. The literature review revealed that safety and security are broad aims of CV technology, and these align with the SME classified topics. Traffic flow efficiency, another broad aim, includes topics such as traffic signaling, information management, vehicle monitoring, vehicle navigation, smart parking, and platooning. Topics focused on wireless communications and allocating vehicular computing resources enable the efficient deployment of CVs. The notable concentration in cybersecurity-related patents indicates a trend toward prioritizing security in CV development. The SME observed that every topic had overlapping objectives. For example, advancements in traffic signaling and multi-vehicle networking can improve both driving safety and the efficiency of traffic flows. Therefore, the SME assigned the dominant topic to each patent summary.

Table 1. SME classification of general patent objectives.

Topic	SME Description
Computing Resource	Communications traffic to exchange sensor data and a wide range of other information, including the need for low latency to meet real-time demands, place additional burden on available computational resources. Objectives target optimal resource allocation and usage of onboard and cloud-based computing resources and optimize communications across multiple network interfaces and servers.
Cybersecurity	Growing wireless connectivity between vehicles and other things, including other vehicles, expands the vulnerability surface for cyber-attacks. Objectives address enhanced cybersecurity, including encryption, authentication, and intrusion detection methods.
Driving Safety	Objectives utilize vehicle-to-everything connectivity and sensors on other vehicles to enhance visibility and situational awareness, safely navigating in complex environments, including through intersections and among pedestrians, and avoiding collisions.
Information Management	Demand for efficient management of information across software applications and services scales with increased vehicle connections. Objectives ensure that systems present relevant information to vehicle operating and in-cabin infotainment systems to prevent data overload and prioritize information that is essential for vehicle operation, safety, and user experience.
Multi-vehicle Networking	Vehicle clusters can form and maintain microvehicular clouds to efficiently share and exchange information. Objectives address the efficient use of resources among vehicles to enable capabilities such as distributed data storage, collaborative computing, reliable communications, and service provisioning.
Platooning	The streamlined aerodynamics resulting from vehicles following each other more closely than normal (platooning) results in better fuel efficiency and improved traffic flow. Objectives address various ways to utilize wireless, sensors, and real-time control mechanisms to enable safer and more cost-efficient platooning and to alert law enforcement.
Smart Parking	Locating parking spaces in crowded and complex environments can be challenging and contribute to congestion. Objectives facilitate cooperative parking space searches, including charging for the "ego" vehicle by using sensors and microvehicular clouds or centralized services.
Traffic Signaling	Suboptimal traffic signal timing can exacerbate congestion. Objectives leverage wireless communications and sensors among vehicles to assess conditions and predict arrival times while dynamically optimizing traffic signaling for overall traffic impact.

Table 1. *Cont.*

Topic	SME Description
Vehicle Monitoring	Objectives aim to enrich in-cabin experiences for passengers through display devices that provide various forms of information and entertainment, and via methods of preventing motion sickness by monitoring and predicting ride quality.
Vehicle Navigation	Objectives are to update electronic maps with real-time data from vehicles for more accurate navigation, and dynamically detecting environmental changes, including topography, emergency situations, and seasonal conditions such as flooding or snow, to inform about alternative routes.
Wireless Communications	Objectives address advancements in wireless communications such as lower latency cellular networks, quality of service, resilience in noisy environments, and interference.

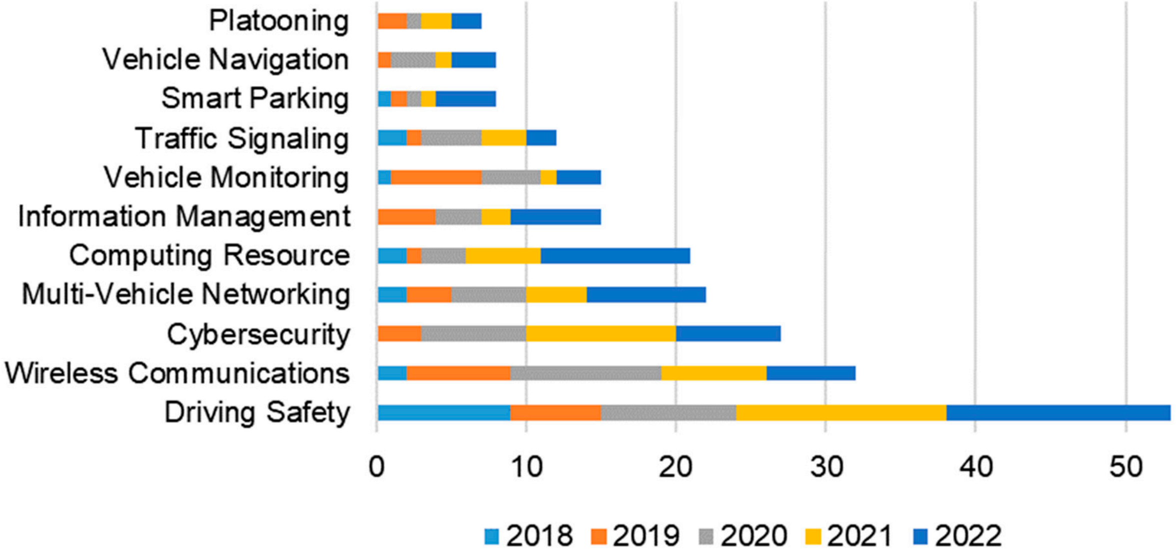

Figure 1. Distribution of patents by category and year issued.

The topic modeling algorithms require cleaned text to provide the best results. Hence, the text cleaning procedure removed noisy text, including characters and words with little or no meaning. These included punctuation marks, numbers, and short tokens such as those containing at most three characters. The field of NLP defines stop words as those that contribute little or no meaning towards understanding a sentence. Common English stop words are *they*, *which*, *that*, *were*, *having*, *doing*, *while*, and *from*. Initially removing short tokens such as *the*, *a*, *an*, *in*, *on*, *at*, or *and* reduces the size of the corpus to speed up the removal of longer stop words, including words that are specific to patent documentation. For example, words commonly used in patent documents are *system*, *method*, *embodiment*, *invention*, *disclosure*, *claim*, *provide*, *describe*, and *include*. In this analysis, the words *connected*, *vehicle*, and *vehicles* were stop words because they must appear in all documents, by definition of the search keywords.

The word cloud of Figure 2 provides a visualization of the important terms in the cleaned corpus of documents within each of the SME-defined categories summarized in Table 1. The word cloud provides a visual confirmation that the SME classification was succinct, rational, understandable, and compelling. A word cloud scales the font size of terms in proportion to their importance in the document set. Terms can be single or multiple adjacent words such as bigrams. The measure of importance is either the term frequency (TF) or a modification of TF, as discussed in the next section. As an example, the word cloud associated with the cybersecurity category included dominant phrases such as *pseudonym certificate*, *authentication*, and *registered master*, which are all related to methods that the patents describe to secure vehicular communications. Similarly, the word cloud for traffic signaling included dominant phrases such as *compatible movement*, *candidate timing*, *entire queue*, and *queue length*, which are all related to controlling traffic lights to manage intersection traffic and queues.

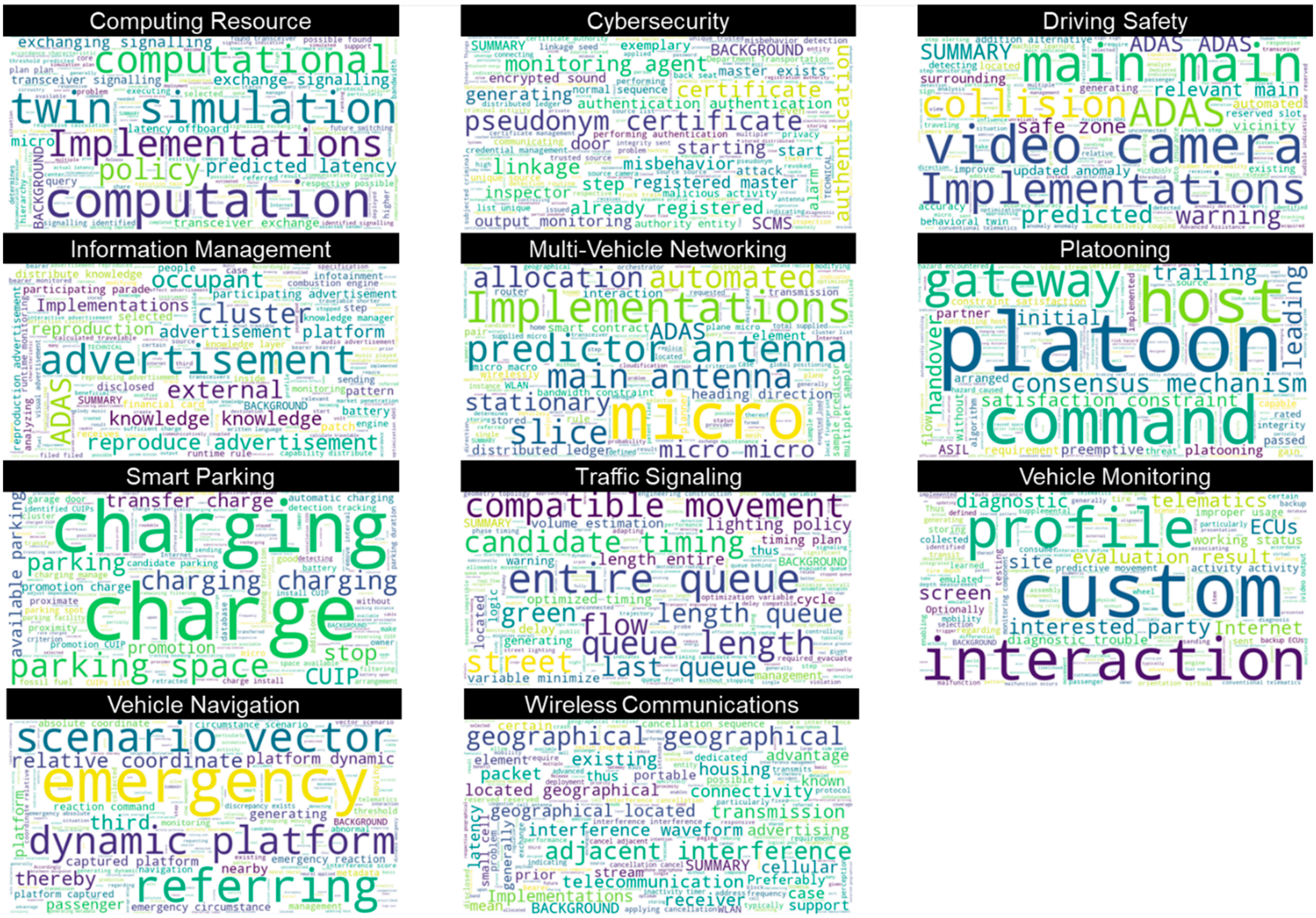

Figure 2. Word cloud of the corpus cleaned by SME classification.

3.2. NLP Pipeline

The NLP pipeline to identify topics started with text tokenization followed by text normalization, text cleaning, vectorization, and an ML algorithm known as topic modeling [37]. Text tokenization converted raw text into individual word fragments called tokens. Text normalization reduced tokens to their base or root form with the aim of minimizing the number of variations that can serve as unique features in the topic modeling process. Text normalizers can be stemmers or lemmatizers. The former truncates the ends of words to a root form, whereas the latter employs a more sophisticated method to identify a base form that accounts for the morphological structure of a word. Therefore, stemmers can overgeneralize words, whereas lemmatizers can retain semantic accuracy at the expense of requiring more computation.

Vectorizers represent the entire corpus as a matrix of values that reflect the importance of words in a document. Each vectorizer has an option to exclude tokens that are too common or too rare. Common thresholds for "too common" and "too rare" are 95% and 5% of the documents, respectively. A vectorizer builds a predetermined size vocabulary of unique tokens in the corpus. The vectorizer output is a sparse matrix of dimensions (N, M), where N is the number of documents in the corpus (rows), and M is the size of the vocabulary (columns). Hence, each entry in the matrix stores the importance of a token (column) within a document (row).

Table 2 summarizes the variety of NLP algorithms available for each part of the NLP pipeline, including their advantages and disadvantages. Lane et al. (2019) provides further in-depth descriptions of these algorithms, which are currently state-of-the-art and mature [37]. The algorithms listed are available in software libraries such as the natural language tool kit (NLTK), scikit-learn (sklearn), and regular expression operations (re) supported by the Python programming language. The NLP libraries constrain tokenizers and normalizers to operate as compatible pairs. For example, the SpaCy tokenizer works only with the SpaCy lemmatizer, but not with any of the other normalizer algorithms.

Table 2. Algorithm options for the NLP pipeline.

	Algorithm	Brief Description	Advantages	Disadvantages
Tokenizer	Whitespace	Splits tokens based on whitespace.	Simple, fast.	Not suitable for languages where whitespace does not denote word boundaries.
	Word	Splits text into words using NLTK's word_tokenize.	Robust, handles punctuation.	Slower compared to whitespace tokenizer.
	Punke	Language-independent tokenizer.	Good for European languages, handles punctuation.	May not work well for languages with different sentence structures.
	Regexp	Tokenization based on regular expression patterns.	Highly customizable.	Requires good understanding of regular expressions.
	Sentence	Splits text into sentences.	Useful for document summarization and segmentation.	Not useful for word-level analysis.
	SpaCy	Tokenizer provided by the SpaCy library.	Fast, handles multiple languages, robust.	Requires installing the SpaCy library and language models.
Normalizer	PorterStemmer	A well-known stemming algorithm.	Effective for English, reduces words to base form.	Over-stemming or under-stemming possible.
	Lancaster Stemmer	More aggressive than Porter.	Reduces words to a basic form.	Can produce stems that are not meaningful words.
	Snowball Stemmer	Extends Porter to support multiple languages.	Language support, aggressive.	Over-stemming or under-stemming possible.
	WordNet Lemmatizer	Lemmatizer based on WordNet lexical database.	Produces valid words, less aggressive than stemming.	Slower, may require POS tags for accurate lemmatization.
	SpaCy Lemmatizer	Lemmatizer offered by the SpaCy library.	Fast, handles multiple languages, usually more accurate.	Requires installing the SpaCy library and language models.
Vectorizer	CountVectorizer	Converts text to a matrix of token counts.	Simple, effective for most cases.	Does not consider semantic meaning, sensitive to frequent words.
	TfidfVectorizer	Converts text to a matrix of TF-IDF features.	Considers global importance of words, less sensitive to frequent words.	Sensitive to the scale of the dataset, more computationally intensive than CountVectorizer.
Topic Model	Latent Dirichlet Allocation	Most used topic modeling algorithm.	Effective for a wide range of topics, easy to interpret.	Requires choosing the number of topics a priori, may not be suitable for all types of text data.
	Non-negative Matrix Factorization	Useful for topic modeling and other types of clustering.	Fast, easier to interpret than LDA.	Assumes linear structure, may not work well for all types of data.
	Latent Semantic Analysis	Also known as latent semantic indexing (LSI).	Good for capturing semantic meaning, less sensitive to word frequency.	High computational cost for large datasets.

The CountVectorizer calculated TF as the feature importance, which is the number of times that a term appeared in a document. The TfidfVectorizer tampers the TF by an inverse document frequency (IDF), effectively attenuating the importance of a word based on its global frequency in the corpus. Hence, if a word appears very frequently in the target document but also frequently in other documents, its importance for that document will be lower.

The unsupervised learning approach to topic modeling offers a scalable, efficient, and practical solution for analyzing large patent datasets. Such models can uncover the underlying thematic structure of a corpus. This contrasts with supervised learning methods that require extensive, accurately labeled data, which is often a challenging and resource-intensive task, especially in complex domains such as patent analysis. Unsupervised learning models effectively cluster the key features of the vectorized corpus. The NLP pipeline focused on evaluating the most mature and best performing topic modeling algorithms reported in the literature. For example, Garbhapu and Bodapati (2020) found that LDA was better than LSA at aligning word associations with human word associations [38]. Kherwa and Bansal (2019) assessed that LDA was more efficient than other topic modeling algorithms [28]. Rüdiger et al. (2022) found that LDA and non-negative matrix factorization (NMF) outperformed most, with LDA better suited for a corpus containing many topic clusters [33].

LDA is a probabilistic model that assumes each document contains a mixture of topics, each represented by probability distributions over a vocabulary of words. LDA iteratively assigns words to topics and updates the topic distributions until it converges to a stable solution. NMF is a deterministic matrix decomposition technique that factors the vectorized corpus into two non-negative matrices. One matrix represents the topics, and the other matrix represents the distribution of topics in each document. NMF finds the factors that minimize the reconstruction error between the original matrix and the product of the factored matrices [37].

3.3. Topic Alignment Index

This research invents a topic alignment index (H-index) to quantify the alignment between NLP topic model assignments and SME classifications. The proposed H-index (H stands for harmonic) comprises two parts. The first part is a *purity* factor defined as the proportion of documents within an SME category represented by the model's most frequently assigned topic in that category. Hence, the average purity factor ρ across C SME-assigned categories is

$$\rho = \frac{1}{C}\sum_{i=1}^{C}\frac{F_i^*}{D_i} \tag{1}$$

where F_i^* is the count of the model's most frequently assigned topic for SME category i and D_i is the number of documents in SME category i. The maximum purity factor is unity when the most frequently assigned topic in an SME category covers all documents in that category, making it a pure assignment. The minimum purity factor is $1/C$ where $F_i^* = 1$, reflecting that no two assigned topics in an SME category were identical. The average purity factor mirrors a homogeneity measure of clustering, which is a global measure of the number of predicted classes that belong to the same cluster [39].

The second part of the proposed H-index is a dispersion factor d that measures the number of most frequently assigned topics to all SME categories that are unique. Hence, dispersion is

$$d = \frac{1}{C}|\boldsymbol{F}| \tag{2}$$

The function $|\cdot|$ denotes the cardinality or count of unique values in the set $\boldsymbol{F} = \{F_1^* \ldots F_C^*\}$ of most frequently assigned topics for SME categories 1 through C. Hence, dispersion measures how well model-assigned topics spread across the SME categories to distinguish among them. The maximum dispersion factor is unity when no two most frequent topic assignments across the SME categories are identical. The minimum dispersion factor is $1/C$ if the model assigned the same topic to all SME categories. Hence, the dispersion factor penalizes repeated topic assignments across the unique SME categories.

The topic alignment index is a combination of the average purity across SME-assigned topics and the dispersion index. A simple combination of the two indices would be the mean value. Alternatively, this research defines the H-index, denoted ψ, as the harmonic mean of the two indices, thereby reflecting a balanced contribution such that

$$\psi = 2\frac{\rho \times d}{\rho + d}. \tag{3}$$

Unlike a simple mean, the harmonic mean captures the influence of outliers. For example, if the average purity factor is unity but the dispersion factor approaches zero, the H-index will also approach zero, whereas the mean value of the two components will instead approach 0.5. This situation represents a scenario where the model assigned the same topic to all SME categories. The situation is the same for the other extreme scenario where the average purity factor approaches zero and the dispersion factor is unity. This situation represents a case where the model assigned a unique topic to each document in an SME category and does so for every SME category. Both cases represent the worst possible alignment with an SME classification, representing the worst possible topic modeling performance.

3.4. Model Optimization

The best performing model identified unique words in the corpus that optimally distinguish among distinct topics. However, stop words manifest as noise that reduces the separability of term distributions among document clusters, which effectively decrease the performance of topic modeling. This research invented a method of stop word engineering to reduce noise, beyond removing common English stop words.

Figure 3 illustrates the two-part workflow to conduct stop word engineering and find the best mix of algorithms in the NLP pipeline. The workflow began by setting the hyperparameter for the number of topic model topics to identify the same number of SME-identified topics and initializing the stop word list to the database search terms. Subsequently, the workflow iteratively selected from an exhaustive combinatorial mix of NLP algorithms in the pipeline {tokenizer, normalizer, vectorizer, filter, topic modeler} to calculate the H-index. The workflow then selected the NLP pipeline that maximized H-index, and then iteratively updated the stop word list. The selected stop words were those that repeated more than once (appeared more than twice) in the combined list of top 20 keywords that the model identified for each topic. The iteration terminated when no keywords repeated more than once. The workflow then iterated through the NLP pipeline again, using the updated stop word list, to identify the best performing pipeline. The pipeline selection terminated once the best NLP pipeline stayed the same.

The rationale for setting the termination criteria to a maximum of a single keyword repetition allowed for each document to reflect a mixture of topics, albeit with one topic dominating.

The rationale for setting the stop word keyword count window to the top 20 was that preliminary analysis established it as a point of diminishing returns in term importance. The Results section plots the keyword frequency distribution of the best model to illustrate and validate this observation. That is, the dominant keywords contribute most toward identifying latent topics, whereas infrequent keywords manifest as noise.

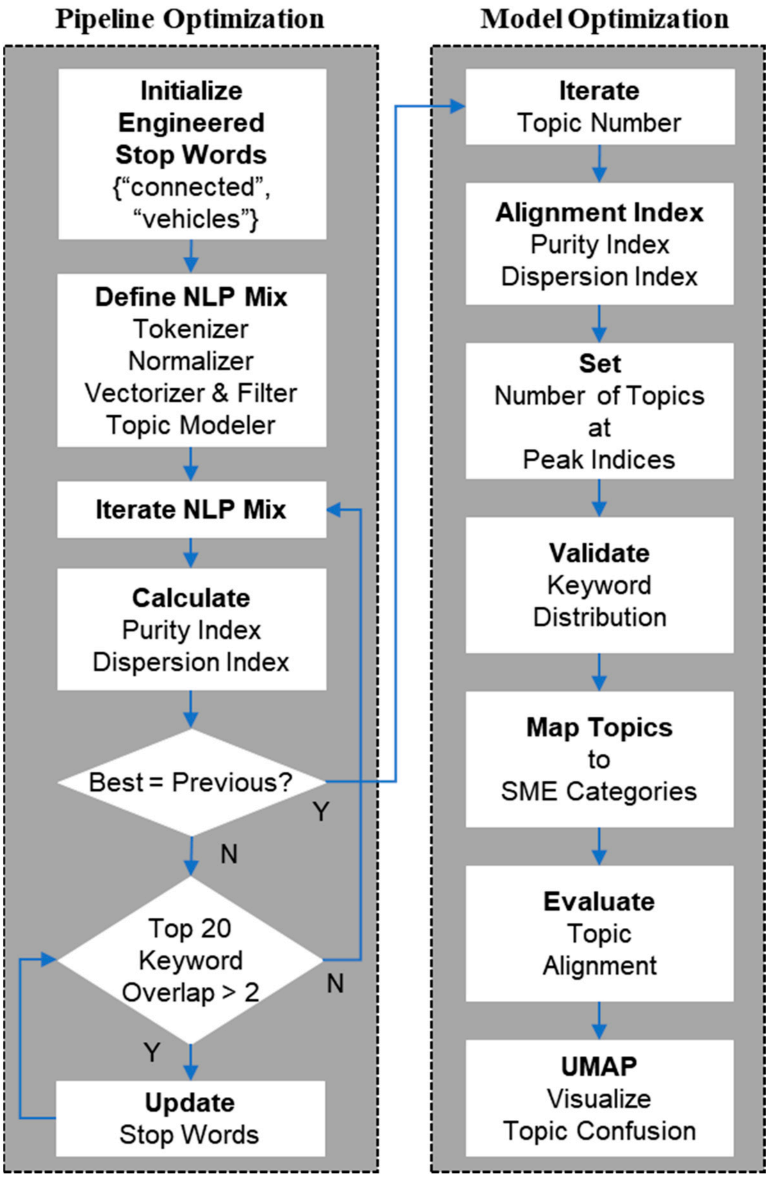

Figure 3. Topic modeling optimization and performance evaluation workflow.

The literature established that topic models share the same flaw in requiring a user to define the topic count for modeling [32]. Hence, the model optimization part of the workflow used the best NLP pipeline from the pipeline to iterate the number of topics to model and calculate the purity, dispersion, and H-indices. The workflow then evaluated topic alignment by selecting the model that provided the peak H-index. The SME then mapped model-assigned topics to the SME categories and empirically assessed a level of qualitative alignment. The workflow then utilized uniform manifold approximation and projection (UMAP) to visualize document clustering by topic assignment probability in a dimensionally reduced feature space, and subsequently to visualize the level of topic assignment confusion.

4. Results

Subsequent subsections describe the process of finding the best NLP pipeline, topic model optimization to find the optimum number of topics, and an assessment of the confusion in topic model assignments relative to the SME-assigned categories.

4.1. Pipeline Optimization

Table 3 summarizes the outcome of the stop word engineering process to identify the best NLP pipeline for model optimization. The process identified 179 English stop words and 372 domain-specific stop words, resulting in 3058 unique tokens in the cleaned corpus, which was approximately 35% of the unique tokens in the original corpus. Vectorization with lower and upper outlier filtering for too frequent and too rare words further reduced the number of tokens (features) to 612, which was only about 7% of the unique tokens in the original corpus. This result highlights the efficiency of the text cleaning process to produce a minimal set of features for ML.

Table 3. Outcome of stop word engineering.

Text Cleaning	Tokens
Unique tokens in the original corpus	8872
Number of English stop words	179
Number of engineered stop words	372
Unique tokens in cleaned corpus	3058
Number of filtered tokens in vectorized dictionary	612

Table 4 summarizes the performance of each NLP pipeline in terms of their alignment index components and the H-index. There were four combinations of tokenizers (T) and normalizers based on the best algorithms identified in Table 2. The NLP pipeline included a mix of the CountVectorizer measuring term frequency (TF) and the TfidfVectorizer measuring TF-IDF (TI), each with options for no filtering (0F) and dual filtering (2F) of extreme importance values. The NLP pipeline included combinations of the LDA and NMF topic modelers. Hence, there were 4 (tokenizer/normalizer) × 2 (vectorizers) × 2 (filters) × 2 (topic modelers) = 32 combinations of NLP algorithms in the NLP pipeline. The results show that NLP pipeline 31 yielded the best H-index.

Table 4. NLP pipelines for 11 topics and alignment indices.

Pipe	Tokenize & Normalize	TF	TI	0F	2F	LDA	NMF	Purity	Dispersion	Mean	H-Index
1	Spacy T & Spacy L	x		x		x		0.386	0.091	0.238	0.147
2	Word T & Word L	x		x		x		0.391	0.182	0.286	0.248
3	Word T & Snowball S	x		x		x		0.443	0.091	0.267	0.151
4	Word T & Porter S	x		x		x		0.462	0.091	0.276	0.152
5	Spacy T & Spacy L	x			x	x		0.347	0.545	0.446	0.424
6	Word T & Word L	x			x	x		0.323	0.636	0.480	0.429
7	Word T & Snowball S	x			x	x		0.349	0.727	0.636	0.472
8	Word T & Porter S	x			x	x		0.365	0.727	0.546	0.486
9	Spacy T & Spacy L		x	x		x		0.392	0.091	0.241	0.148
10	Word T & Word L		x	x		x		0.428	0.182	0.305	0.255
11	Word T & Snowball S		x	x		x		0.442	0.091	0.267	0.151
12	Word T & Porter S		x	x		x		0.452	0.091	0.271	0.151
13	Spacy T & Spacy L		x		x	x		1.000	0.091	0.545	0.167
14	Word T & Word L		x		x	x		0.641	0.182	0.412	0.283
15	Word T & Snowball S		x		x	x		0.904	0.091	0.498	0.165
16	Word T & Porter S		x		x	x		0.953	0.091	0.522	0.166
17	Spacy T & Spacy L	x		x			x	0.372	0.182	0.277	0.244
18	Word T & Word L	x		x			x	0.378	0.182	0.280	0.245
19	Word T & Snowball S	x		x			x	0.403	0.182	0.292	0.250
20	Word T & Porter S	x		x			x	0.411	0.182	0.297	0.252
21	Spacy T & Spacy L	x			x		x	0.728	0.182	0.455	0.291
22	Word T & Word L	x			x		x	0.585	0.182	0.383	0.277
23	Word T & Snowball S	x			x		x	0.537	0.364	0.450	0.434

Table 4. *Cont.*

Pipe	Tokenize & Normalize	TF	TI	0F	2F	LDA	NMF	Purity	Dispersion	Mean	H-Index
24	Word T & Porter S	x			x		x	0.696	0.182	0.439	0.288
25	Spacy T & Spacy L		x	x			x	0.443	0.091	0.267	0.151
26	Word T & Word L		x	x			x	0.441	0.091	0.266	0.151
27	Word T & Snowball S		x	x			x	0.433	0.091	0.262	0.150
28	Word T & Porter S		x	x			x	0.468	0.091	0.279	0.152
29	Spacy T & Spacy L		x		x		x	0.493	0.455	0.474	0.473
30	Word T & Word L		x		x		x	0.522	0.545	0.534	0.533
31	Word T & Snowball S		x		x		x	0.524	0.545	0.535	0.535
32	Word T & Porter S		x		x		x	0.483	0.545	0.514	0.512

Figure 4 is a visualization of the data in Table 4 to show trends. The figure shows the NLP pipelines in eight groups. The following are some observations and insights:

1. The vectorizer extremity filters improved the performance of both LDA (group 2 > group 1) and NMF (group 8 > group 7) topic modeling algorithms.
2. LDA had an edge over NMF when the pipeline included TF and 2F (group 2 > group 6).
3. NMF had an edge over LDA when the pipeline included TI and 2F (group 8 > group 4).
4. In all cases, stemmers performed at least as good as the lemmatizers.

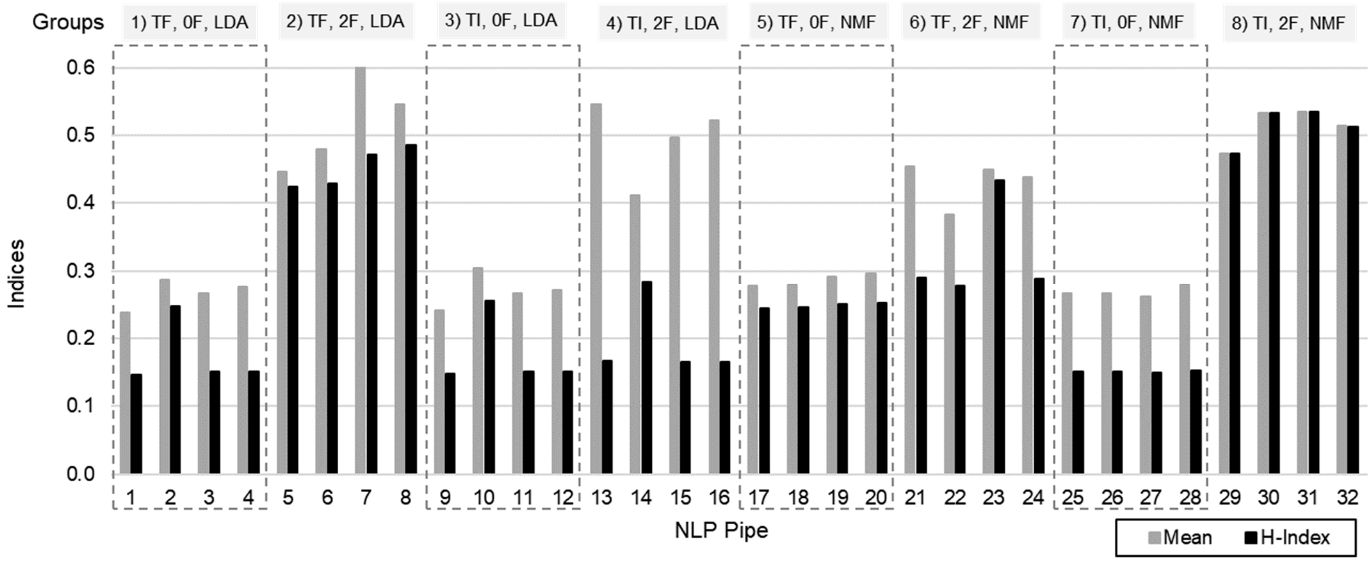

Figure 4. Alignment indices of each NLP pipeline.

Table 5 lists the top 20 keywords associated with each of the topics (T) assigned by the best NLP pipeline. Figure 5 shows the keyword distribution by TF-IDF for each topic that the model assigned. This distribution validates that there is a point of diminishing returns in keyword importance after 5 to 20 top keywords, depending on the model-assigned topic. For example, topics 2 and 8 exhibit a point of diminishing returns after approximately 5 keywords, whereas topics 3 and 6 show that diminishing returns become apparent as the number of keywords approaches 20.

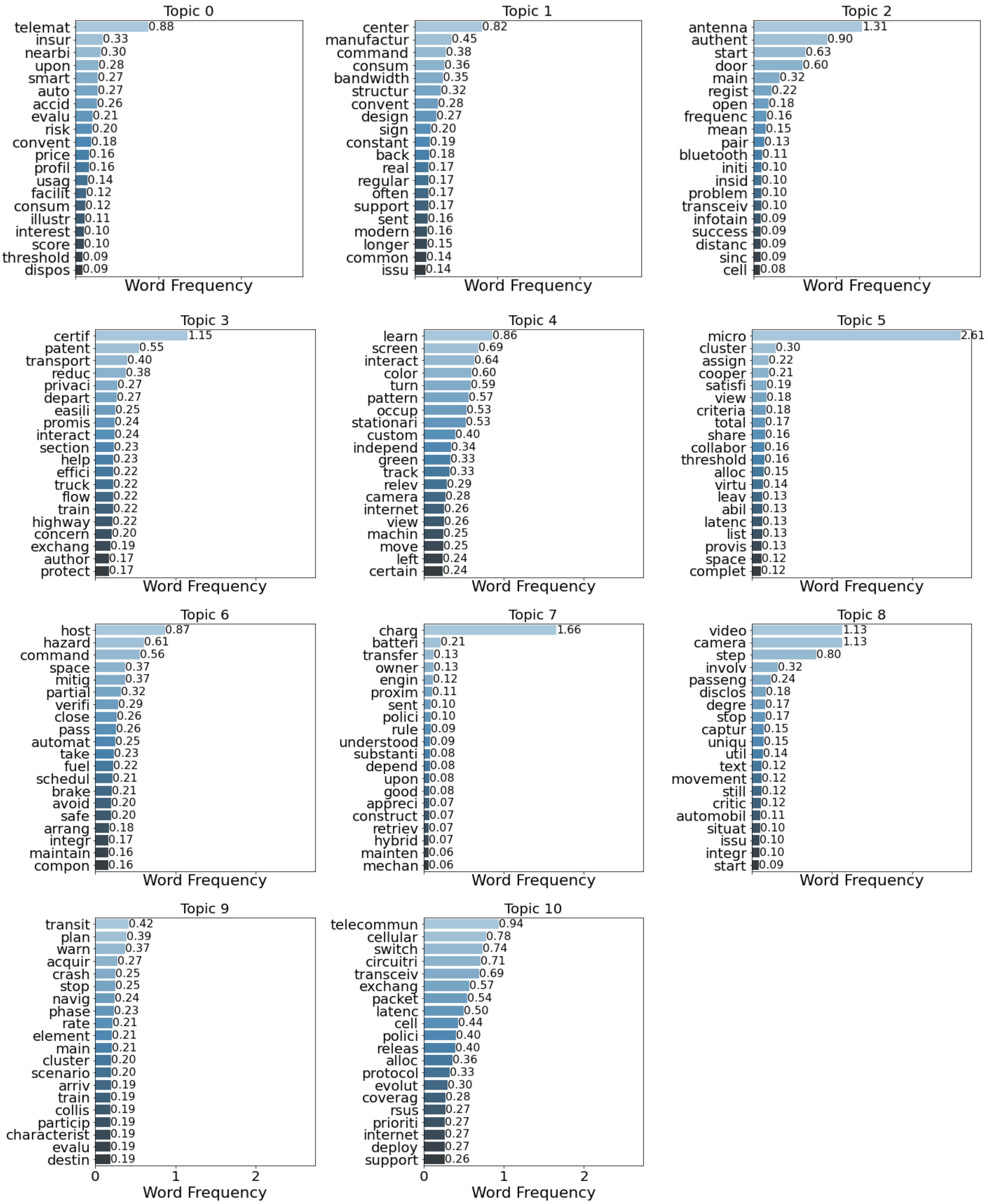

Figure 5. Term TF-IDF distribution per assigned topic category.

Table 5. Topic number and associated top 20 keywords.

T	Top 20 Keywords
0	telemat insur nearbi upon smart auto accid evalu risk convent price profil usag facilit consum illustr interest score threshold dispos
1	center manufactur command consum bandwidth structur convent design sign constant back real regular often support sent modern longer common issu
2	antenna authent start door main regist open frequenc mean pair bluetooth initi insid problem transceiv infotain success distanc sinc cell
3	certif patent transport reduc privaci depart easili promis interact section help effici truck flow train highway concern exchang author protect
4	learn screen interact color turn pattern occup stationari custom independ green track relev camera internet view machin move left certain
5	micro cluster assign cooper satisfi view criteria total share collabor threshold alloc virtu leav abil latenc list provis space complet
6	host hazard command space mitig partial verifi close pass automat take fuel schedul brake avoid safe arrang integr maintain compon
7	charg batteri transfer owner engin proxim sent polici rule understood substanti depend upon good appreci construct retriev hybrid mainten mechan
8	video camera step involv passeng disclos degre stop captur uniqu util text movement still critic automobil situat issu integr start
9	transit plan warn acquir crash stop navig phase rate element main cluster scenario arriv train collis particip characterist evalu destin
10	telecommun cellular switch circuitri transceiv exchang packet latenc cell polici releas alloc protocol evolut coverag rsus prioriti internet deploy support

Figure 6 complements Table 5 and Figure 5 by showing a word cloud for each of the model-assigned topics. To improve visualization, the font sizes correspond to the TF of the keyword distributions rather than the TI-IDF values. The visualization shows both distinct and overlapping topics. For example, topic 6 is clearly about vehicle platooning, whereas topic 9 includes a mix of topics based on words such as transit, warn, crash, navigate, and cluster.

Table 6 shows the topic number most frequently assigned to each SME category and the purity value. The ML model assigned topic 9 to the SME category of traffic signaling with maximum purity. This aligns with the top keywords of topic 9, which include transit, plan, warn, crash, stop, navigate, and phase. However, the model also assigned topic 9 to five other SME categories, resulting in a low dispersion of 6/11 = 0.545. The minimum purity of 0.318 for the SME category of multi-vehicle networking suggested that the topic model identified several topics in that category. Nevertheless, assigning topic 5 as the dominant topic aligns with multi-vehicle networking, where the top keywords include micro, cluster, cooperative, and share. The average purity for the best NLP pipeline was 0.524, resulting in both the mean and H-index equaling 0.535.

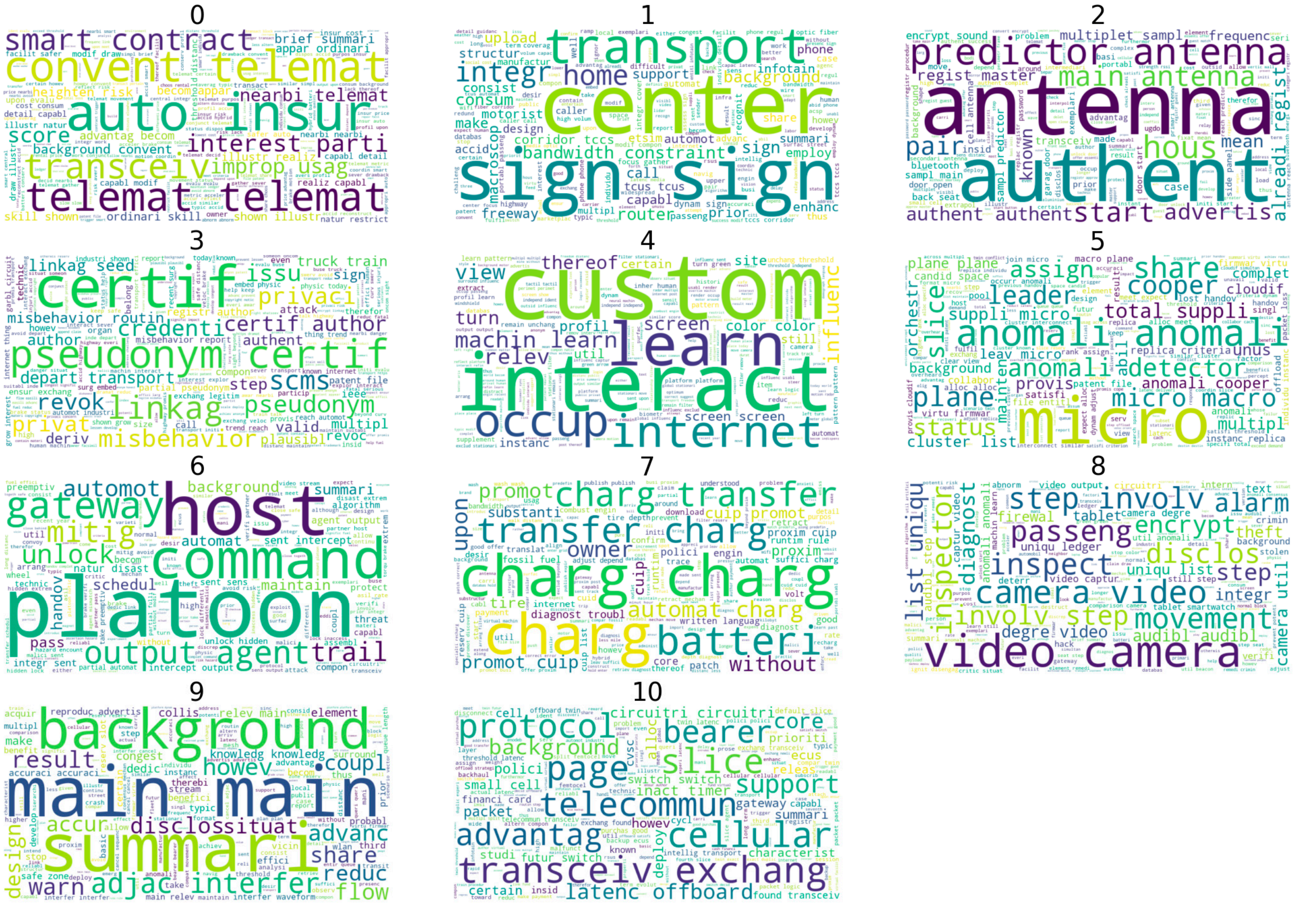

Figure 6. Word cloud for topic assignments using the top NLP pipeline.

Table 6. Most frequent topics assigned to each SME class.

SME Category	Most Frequent Topic	Purity
Computing Resource	10	0.333
Cybersecurity	3	0.444
Driving Safety	9	0.472
Information Management	9	0.533
Multi-Vehicle Networking	5	0.318
Platooning	6	0.857
Smart Parking	7	0.375
Traffic Signaling	9	1.000
Vehicle Monitoring	9	0.400
Vehicle Navigation	9	0.625
Wireless Communications	9	0.406

4.2. Model Optimization

Both topic modeling algorithms (LDA and NMF) require a hyperparameter that specifies the number of topics to model. Figure 7 shows trends of how the various indices change with the number of topics modeled for the best NLP pipeline. The H-index peaked at 11 topics, aligning with the number of SME-assigned categories. This result affirmed that the best model aligned with the SME classification, albeit with a fair amount of disagreement in topic assignments. The purity and dispersion factors show diverging trends, as expected. That is, as the dispersion of topic assignments increases within and across SME categories, the purity of topic assignments within SME categories will necessarily decrease. Conversely, as the purity of topic assignments within and across SME categories increases, the dispersion factor must necessarily decrease. Unlike the H-index, the extreme values of purity and dispersion factors more heavily influenced their mean value. After peaking at 11 topics, the H-index tended to stabilize with an increasing number of topics modeled.

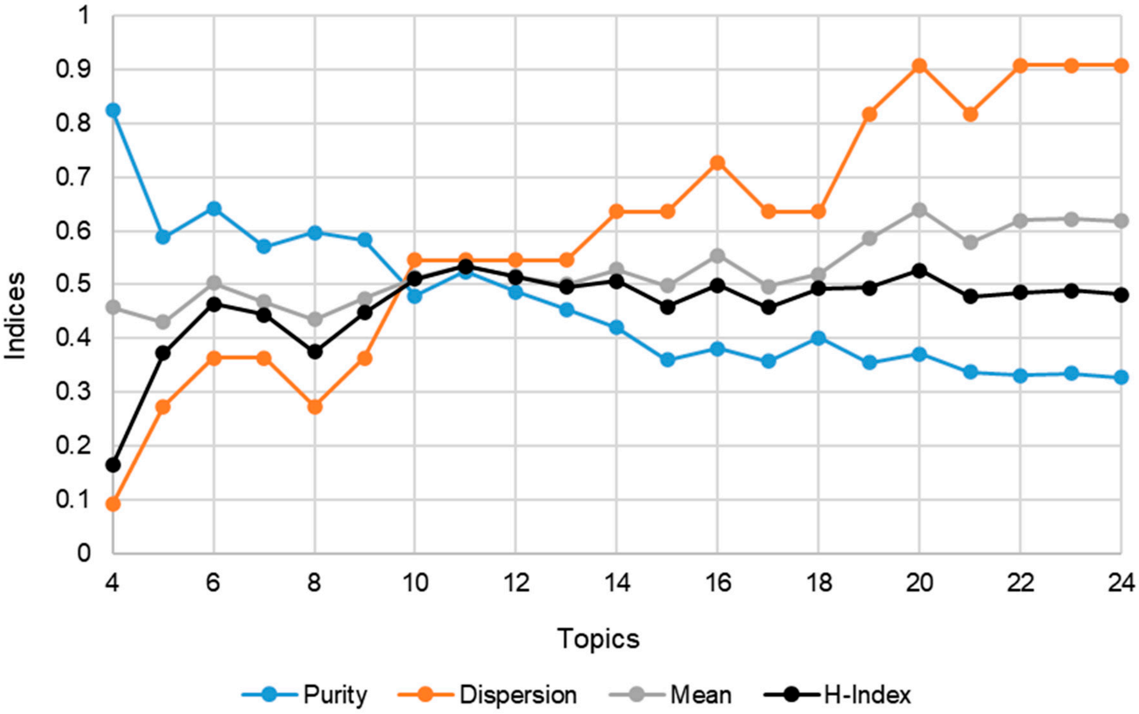

Figure 7. Topic model hyperparameter optimization.

4.3. Misalignment Visualization

Table 7 offers some insights into the amount of misalignment in model topic assignments with the SME-assigned categories. The table shows that topics (T) assigned by the model align with only 6 of the 11 SME-assigned categories. Most topics assigned by the model related to the SME category of driving safety. The representative keywords from the top 20 list, and as visualized in the word cloud (Figure 6), offer a glimpse into the reason for the topic assignment. For instance, the model assigned topic 1 to documents that contained *sign* as the dominant keyword. Those documents were mostly within the SME-assigned category of driving safety. However, as shown in the table, the model assigned different topics to documents that also pertained to the SME classification of driving safety. The table also lists the number of unique SME categories (USC) contained within each of the model-defined topic categories. Topic 3 achieved the best alignment with the SME category of cybersecurity with a single unique assignment. Topic 9 achieved the worst alignment because it contained all 11 SME categories.

The topic modeling output is a matrix of documents (rows) and topic probabilities (columns) that represent the mix of all topics in a document. However, the topic assigned to a document is the one with the highest probability. Similar probability vectors of topic assignments will cluster in feature space. However, it is not possible nor practical to visualize clusters of documents, each with 11 topic probabilities or features. Fortunately, uniform manifold approximation and projection (UMAP) is a technique used to visualize high-dimensional data in two-dimensional space by embedding features in a manner that preserves their original structure and relationships [40].

Table 7. Topic model confusion with SME categories.

T	Majority SME Category	USC	Topic Model Keywords Relevant to SME Category
0	Driving Safety	4	insurance, accident, risk
1	Driving Safety	5	sign
2	Cybersecurity	4	authenticate, register
3	Cybersecurity	1	certificate, privacy
4	Driving Safety	5	green, camera, occupy, interact
5	Multi-Vehicle Networking	4	micro, anomali, slice
6	Platooning	7	platoon, host, gateway, command, trail, brake, automate
7	Smart Parking	4	charge, battery, transfer, promotion
8	Driving Safety	3	video camera, capture, situation, stop
9	Driving Safety	11	interfere, collision, crash, stop, navigation, warn
10	Wireless Communications	6	telecommunication, cellular, transceiver, packet, switch, cell

Figure 8 shows the UMAP of topic model probabilities, labeled by the topic model assignments. It is evident that the document clusters reflect the maximum probability topic assigned. The chart also reveals how documents contain a mix of topics because they form overlapping clusters, such as near the coordinates (7.5, 5.0). As highlighted by the dotted line boundaries shown, clusters such as topic 2, topic 5, and topic 8 are more isolated and homogeneous in topic mix. Figure 9 labels the UMAP of these topic model probabilities with the SME topic assignments to highlight the misalignment between model-assigned and SME-assigned categories. It is clear from the chart that even the remote clusters are no longer homogeneous.

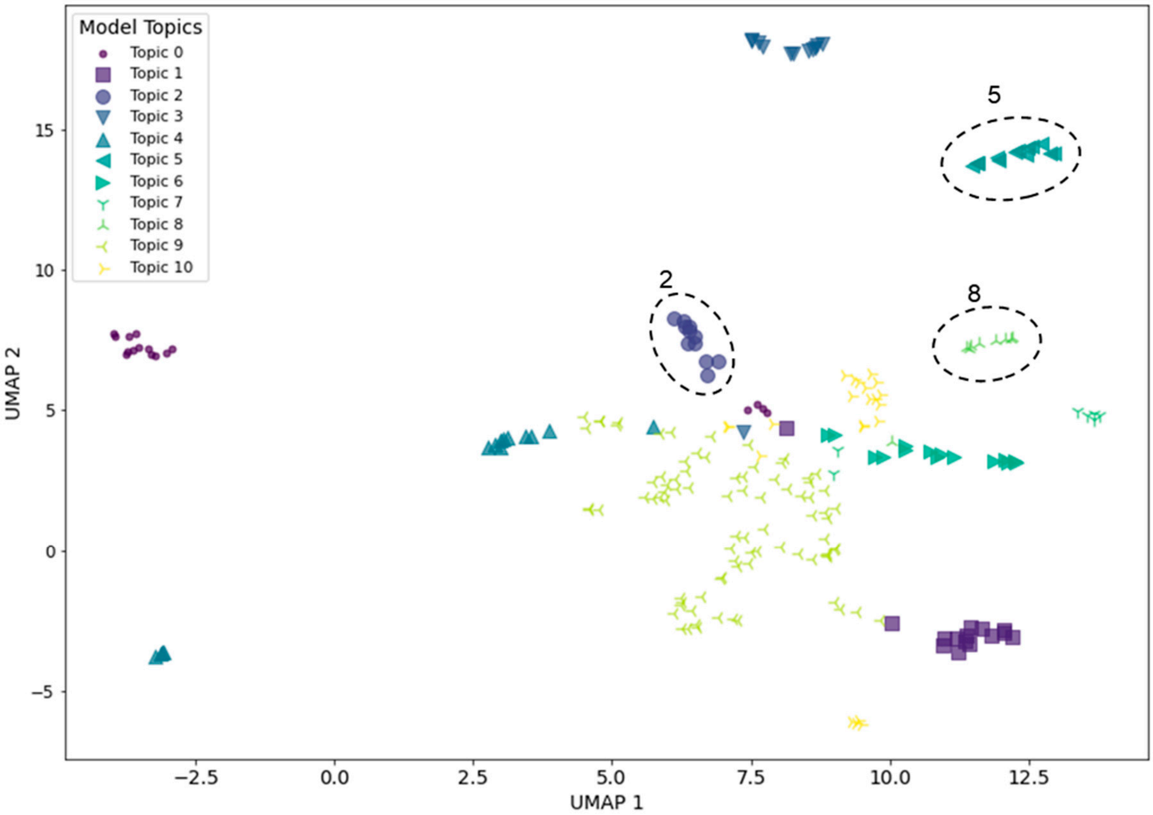

Figure 8. UMAP of topic model probabilities, colored by model topic assignments.

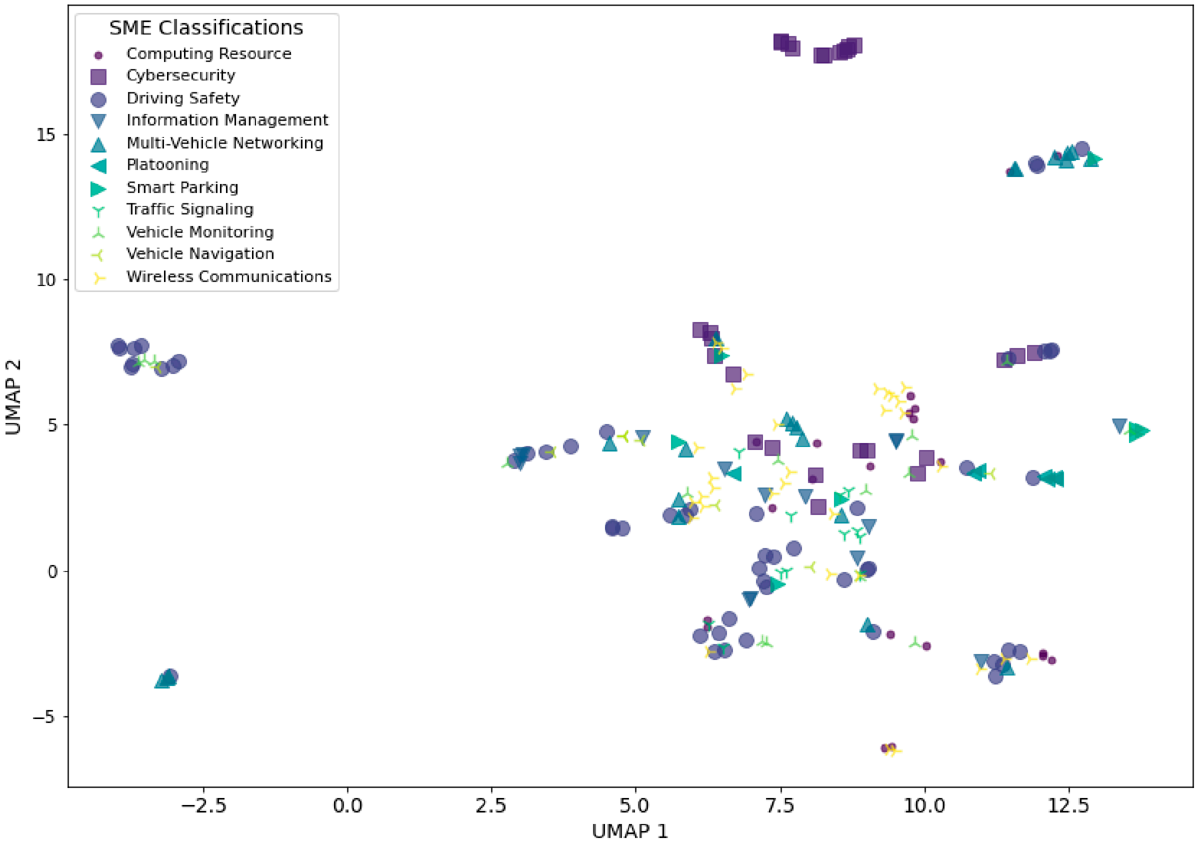

Figure 9. UMAP of topic model probabilities, colored by SME topic assignments.

5. Discussion

This research addressed a critical gap in the automated analysis of complex patent data by developing a new metric to quantify the alignment between subject matter expert (SME) classifications and NLP-based topic assignments. However, this research has a few limitations, which set the stage for future work. Although the author's expertise and experience in the domain of CV is substantial, relying on a single SME to identify dominant themes can result in subjective bias. Nevertheless, the novel metric developed in this work establishes a foundation for future work to investigate the nature of a distribution in the misalignment between SME and NLP topic identification. To do so, future work will select multiple SMEs with different technical backgrounds to identify dominant themes from the same corpus and evaluate the distribution of misalignment with NLP topic classification.

Another limitation is the focus on U.S. patents, which may miss broader international themes. However, foreign patents are not consistently available in the English language, making a mixture analysis more challenging. The results of this work highlight that the gross misalignment of ML-assigned topics with SME classification questions the generalizability of NLP in patent analysis, including across other domains. Therefore, future work will evaluate the same approach on patents focused on related technologies such as autonomous vehicles, advanced air mobility, battery technology, and cybersecurity.

This study established a benchmark in a snapshot of time for future work to compare with by using the metric proposed for objective and quantitative comparison. The evolution from free open-source NLP models, such as the ones evaluated in this research, to LLMs, including more powerful multimodal versions currently under development by companies such as OpenAI, Google, X Corp. (formerly Twitter), Meta, and Microsoft, could also help to identify topics. Hence, future research will include various LLMs in the SME mix to evaluate their misalignment with traditional NLP topic models.

Overall, this work delivers valuable insights to various stakeholders involved in technology development and innovation. Stakeholders include patent analysts and researchers who often face challenges in manually sifting through vast amounts of patent data to identify trends and key areas of innovation. The unsupervised methods can automatically categorize patents into coherent topics, significantly aiding in the quick identification of technological trends and focus areas. Additionally, policy makers and strategic planners can benefit by gaining a macro-level view of technological advancements in the CV domain to inform decision making and strategic planning, especially in technology forecasting and policy formulation. Furthermore, insights gained from topic classification can help innovators and entrepreneurs identify under-explored areas and potential opportunities for innovation, thereby guiding further research and development efforts.

6. Conclusions

This multidisciplinary and interdisciplinary study contributes to the field of natural language processing (NLP) and connected vehicle (CV) technology development. The primary contribution is the development of a novel metric (the H-index) to quantify the alignment between subject matter expert (SME) classifications and topic assignments by state-of-the-art, open-source NLP and ML models. This metric, comprising a purity factor and a dispersion factor, offers a quantitative and objective understanding of the effectiveness of various NLP and ML approaches in capturing the essence of patent documents covering complex technological domains such as CVs.

This research navigated the challenges of analyzing a vast and unstructured dataset of patent documents. The NLP pipeline included a variety of the top performing tokenizers, normalizers, vectorizers, and topic modelers to demonstrate the potential of ML models to categorize and produce insights into the intricate patterns within patent documents. The results highlighted the misalignment of these models with SME classifications by revealing areas where the models diverged, offering critical insights into their limitations and areas for improvement. The integration of manifold visualizations provided a clear and intuitive representation of the areas of misalignment. This visualization tool can help people interpret the complex relationships and patterns within the data, making the findings more accessible to both technical and non-technical stakeholders.

In broader implications, it is evident that while there has been significant progress in the field, challenges remain. The complexity of CV technology, combined with the rapidly evolving landscape of patent information, requires continuous refinement of combined NLP and ML techniques. Future research should assess the interpretability and accuracy of these models in other domains to benchmark their performance for comparison with evolving tools such as large language models (LLMs) and their multimodal variants.

In summary, this study is a step forward in the intersection of NLP, ML, and patent analysis. By developing a metric that bridges the gap between algorithmic efficiency and human-centric interpretability, this research contributes a useful tool for researchers and practitioners in the field. It paves the way for future innovations in NLP and ML and sets a precedent for the rigorous evaluation of these models in the analysis of patent data, particularly in specialized domains such as CV technology.

Funding: This research was funded by the United States' Department of Transportation, grant name Mountain Plains Consortium.

Data Availability Statement: Data are contained within the article.

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

References

1. Casola, S.; Lavelli, A. Summarization, simplification, and generation: The case of patents. *Expert Syst. Appl.* **2022**, *205*, 117627. [CrossRef]
2. Krestel, R.; Chikkamath, R.; Hewel, C.; Risch, J. A survey on deep learning for patent analysis. *World Pat. Inf.* **2021**, *65*, 102035. [CrossRef]

3. Borghesani, V.; Armoza, J.; Hebart, M.N.; Bellec, P.; Brambati, S.M. The Three Terms Task—An open benchmark to compare human and artificial semantic representations. *Sci. Data* **2023**, *10*, 1–13. [CrossRef] [PubMed]
4. USDOT. *Vehicle-to-Everything (V2X) Communications Summit: Detailed Meeting Summary: Preparing for Connected, Interoperable Deployment Nationwide*; United States Department of Transportation (USDOT): Washington, DC, USA, 2023.
5. Nkenyereye, L.; Nkenyereye, L.; Jang, J.-W. Convergence of Software-Defined Vehicular Cloud and 5G Enabling Technologies: A Survey. *Electronics* **2023**, *12*, 2066. [CrossRef]
6. Shichun, Y.; Zheng, Z.; Bin, M.; Yifan, Z.; Sida, Z.; Mingyan, L.; Yu, L.; Qiangwei, L.; Xinan, Z.; Mengyue, Z.; et al. Essential Technics of Cybersecurity for Intelligent Connected Vehicles: Comprehensive Review and Perspective. *IEEE Internet Things J.* **2023**, *10*, 21787–21810. [CrossRef]
7. Rathore, R.S.; Hewage, C.; Kaiwartya, O.; Lloret, J. In-Vehicle Communication Cyber Security: Challenges and Solutions. *Sensors* **2022**, *22*, 6679. [CrossRef] [PubMed]
8. Ju, Z.; Zhang, H.; Li, X.; Chen, X.; Han, J.; Yang, M. A Survey on Attack Detection and Resilience for Connected and Automated Vehicles: From Vehicle Dynamics and Control Perspective. *IEEE Trans. Intell. Veh.* **2022**, *7*, 815–837. [CrossRef]
9. Hildebrand, B.; Baza, M.; Salman, T.; Tabassum, S.; Konatham, B.; Amsaad, F.; Razaque, A. A comprehensive review on blockchains for Internet of Vehicles: Challenges and directions. *Comput. Sci. Rev.* **2023**, *48*, 100547. [CrossRef]
10. Khan, R.; Mehmood, A.; Iqbal, Z.; Maple, C.; Epiphaniou, G. Security and Privacy in Connected Vehicle Cyber Physical System Using Zero Knowledge Succinct Non Interactive Argument of Knowledge over Blockchain. *Appl. Sci.* **2023**, *13*, 1959. [CrossRef]
11. Alanazi, F. A Systematic Literature Review of Autonomous and Connected Vehicles in Traffic Management. *Appl. Sci.* **2023**, *13*, 1789. [CrossRef]
12. Shi, Y.; Wang, Z.; LaClair, T.J.; Wang, C.; Shao, Y. Real-time control of connected vehicles in signalized corridors using pseudospectral convex optimization. *Optim. Control. Appl. Methods* **2023**, *44*, 2257–2277. [CrossRef]
13. Gholamhosseinian, A.; Seitz, J. A Comprehensive Survey on Cooperative Intersection Management for Heterogeneous Connected Vehicles. *IEEE Access* **2022**, *10*, 7937–7972. [CrossRef]
14. Xu, J.; Tian, Z. OD-Based Partition Technique to Improve Arterial Signal Coordination Using Connected Vehicle Data. *Transp. Res. Rec. J. Transp. Res. Board* **2022**, *2677*, 252–265. [CrossRef]
15. Wang, B.; Han, Y.; Wang, S.; Tian, D.; Cai, M.; Liu, M.; Wang, L. A Review of Intelligent Connected Vehicle Cooperative Driving Development. *Mathematics* **2022**, *10*, 3635. [CrossRef]
16. Cui, G.; Zhang, W.; Xiao, Y.; Yao, L.; Fang, Z. Cooperative Perception Technology of Autonomous Driving in the Internet of Vehicles Environment: A Review. *Sensors* **2022**, *22*, 5535. [CrossRef] [PubMed]
17. Gao, B.; Wan, K.; Chen, Q.; Wang, Z.; Li, R.; Jiang, Y.; Mei, R.; Luo, Y.; Li, K. A Review and Outlook on Predictive Cruise Control of Vehicles and Typical Applications Under Cloud Control System. *Mach. Intell. Res.* **2023**, *20*, 614–639. [CrossRef]
18. Islam, Z.; Abdel-Aty, M. Traffic conflict prediction using connected vehicle data. *Anal. Methods Accid. Res.* **2023**, *39*, 100275. [CrossRef]
19. Schwarz, C.; Wang, Z. The Role of Digital Twins in Connected and Automated Vehicles. *IEEE Intell. Transp. Syst. Mag.* **2022**, *14*, 41–51. [CrossRef]
20. Trappey, A.J.; Trappey, C.V.; Wu, J.-L.; Wang, J.W. Intelligent compilation of patent summaries using machine learning and natural language processing techniques. *Adv. Eng. Inform.* **2019**, *43*, 101027. [CrossRef]
21. Joshi, U.; Hedaoo, M.; Fatnani, P.; Bansal, M.; More, V. Patent Classification with Intelligent Keyword Extraction. In Proceedings of the 2022 6th International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 26–27 August 2022; pp. 1–7.
22. De Clercq, D.; Diop, N.-F.; Jain, D.; Tan, B.; Wen, Z. Multi-label classification and interactive NLP-based visualization of electric vehicle patent data. *World Pat. Inf.* **2019**, *58*, 101903. [CrossRef]
23. Hyun, Y.-G.; Han, J.-H.; Chae, U.; Lee, G.-H.; Lee, J.-Y. A study on technical trend analysis related to semantic analysis of NLP through domestic/foreign patent data. *J. Digit. Converg.* **2020**, *18*, 137–146.
24. Wu, H.; Shen, G.; Lin, X.; Li, M.; Zhang, B.; Li, C.Z. Screening patents of ICT in construction using deep learning and NLP techniques. *Eng. Constr. Arch. Manag.* **2020**, *27*, 1891–1912. [CrossRef]
25. Arts, S.; Hou, J.; Gomez, J.C. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Res. Policy* **2021**, *50*, 104144. [CrossRef]
26. Puccetti, G.; Giordano, V.; Spada, I.; Chiarello, F.; Fantoni, G. Technology identification from patent texts: A novel named entity recognition method. *Technol. Forecast. Soc. Chang.* **2023**, *186*, 122160. [CrossRef]
27. de Rezende, J.M.; Rodrigues, I.M.d.C.; Resendo, L.C.; Komati, K.S. Combining natural language processing techniques and algorithms LSA, word2vec and WMD for technological forecasting and similarity analysis in patent documents. *Technol. Anal. Strat. Manag.* **2022**, 1–22. [CrossRef]
28. Kherwa, P.; Bansal, P. Topic modeling: A comprehensive review. *EAI Endorsed Trans. Scalable Inf. Syst.* **2019**, 7. [CrossRef]
29. Abdelrazek, A.; Eid, Y.; Gawish, E.; Medhat, W.; Hassan, A. Topic modeling algorithms and applications: A survey. *Inf. Syst.* **2023**, *112*, 102131. [CrossRef]
30. Meaney, C.; Stukel, T.A.; Austin, P.C.; Moineddin, R.; Greiver, M.; Escobar, M. Quality indices for topic model selection and evaluation: A literature review and case study. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 1–18. [CrossRef]

31. Harrando, I.; Lisena, P.; Troncy, R. Apples to apples: A systematic evaluation of topic models. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Online, 1–3 September 2021.
32. Vayansky, I.; Kumar, S.A.P. A review of topic modeling methods. *Inf. Syst.* **2020**, *94*, 101582. [CrossRef]
33. Rüdiger, M.; Antons, D.; Joshi, A.M.; Salge, T.-O. Topic modeling revisited: New evidence on algorithm performance and quality metrics. *PLoS ONE* **2022**, *17*, e0266325. [CrossRef]
34. Hoyle, A.; Goel, P.; Hian-Cheong, A.; Peskov, D.; Boyd-Graber, J.; Resnik, P. Is automated topic model evaluation broken? The Incoherence of Coherence. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 2018–2033.
35. WIPO. *IP Facts and Figures*; World Intellectual Property Organization (WIPO): Geneva, Switzerland, 2022.
36. USPTO. Data Download Tables. U. P. [USPTO], 20 September 2023. Available online: https://patentsview.org/download/brf_sum_text (accessed on 2 October 2023).
37. Lane, H.; Howard, C.; Hapke, H.M. *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python*; Manning Publications Co., Ltd.: Shelter Island, NY, USA, 2019.
38. Garbhapu, V.K.; Bodapati, P. A comparative analysis of Latent Semantic analysis and Latent Dirichlet allocation topic modeling methods using Bible data. *Indian J. Sci. Technol.* **2020**, *13*, 4474–4482. [CrossRef]
39. Rosenberg, A.; Hirschberg, J. V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 28–30 June 2007.
40. Becht, E.; McInnes, L.; Healy, J.; Dutertre, C.-A.; Kwok, I.W.H.; Ng, L.G.; Ginhoux, F.; Newell, E.W. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **2018**, *37*, 38–44. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.