ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 1 September 2025; Accepted: 13 October 2025

https://doi.org/10.17559/TV-20250901002936
Original scientific paper

BERTopic Modeling Analysis of Privacy Protection Trends in New Trade Rules for the Mobility Industry

Kyoungjin KIM, Haneul HAN*, Sangjo YOO

Abstract: The rapid transformation of the mobility industry has led to an unprecedented volume of personal data, posing new privacy challenges within evolving trade rules. This study applies BERTopic modeling, a BERT-based approach, to systematically analyze 187 screened academic papers (2020-2025) on privacy regulations in the mobility sector. The final corpus was assembled through database-driven filtering (initial n = 321, deduplication, relevance screening). Unlike traditional topic models (e.g., LDA), BERTopic's contextual embeddings facilitate superior semantic understanding, which is crucial for nuanced interdisciplinary analysis. Key hyperparameters include UMAP ($n_{neighbors}$ = 15), HDBSCAN ($min_{cluster}$ = 10), and c-TF-IDF weighting. Through BERTopic analysis, five distinct research themes emerged, health security, user acceptance, location protection, smart city innovation, and autonomous vehicle AI integration, demonstrating its superior performance over LDA in topic coherence (+26.6%), diversity (+8.9%), and keyword accuracy (+19.9%). A comparative policy analysis of GDPR, CCPA, and Korea's PIPA reveals significant regulatory gaps. The findings provide critical insights for policymakers, highlighting the need for harmonized mobility-specific privacy frameworks and offering technical recommendations such as differential privacy for UAM and federated learning for traffic prediction. This study demonstrates that context-aware topic modeling reveals intricate and evolving trends in mobility privacy, addressing the limitations of prior research that lacked the granularity to capture such complexities.

Keywords: BERT; intelligent transportation systems; international trade regulations; mobility data; privacy protection; topic modelling

1 INTRODUCTION

The fourth industrial revolution has catalyzed unprecedented transformation in the transportation sector, driven by the convergence of connectivity, autonomous, sharing, and electrification (CASE) technologies [1]. This evolution encompasses electric vehicles, shared mobility platforms, autonomous driving systems, and Urban Air Mobility (UAM), collectively generating massive volumes of personal data, including location information, movement patterns, vehicle performance metrics, and user behavioural data [2].

The proliferation of mobility data presents dual challenges: maximizing innovation potential while ensuring robust privacy protection across diverse regulatory jurisdictions [3]. Current mobility services routinely collect vehicle identification numbers (VIN), battery identification numbers (BIN), GPS trajectories, charging patterns, and high-definition (HD) map data, all of which carry significant re-identification risks [4]. For instance, electric vehicle operation data can reveal detailed lifestyle patterns when combined with user account information, while shared mobility GPS data retain re-identification potential even after 15-minute aggregation intervals [5].

This rapid evolution has intensified the focus on data privacy, particularly given the fragmented and evolving international regulatory landscape. Understanding the trends in privacy protection amidst new trade rules and diverse regulatory frameworks (such as GDPR, CCPA, and Korea's PIPA) is therefore critical to ensure responsible innovation and cross-border data flow in the mobility sector. This study aims to precisely identify these trends and regulatory challenges through advanced textual analysis.

International privacy regulations, including the European Union's General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and Korea's Personal Information Protection Act (PIPA), establish divergent frameworks for cross-border data transfers, creating friction in global mobility commerce [6]. These regulatory disparities particularly impact emerging mobility technologies requiring real-time international data sharing, such as autonomous vehicle fleets and cross-border UAM services. While existing research has examined specific privacy technologies or individual regulatory frameworks, a comprehensive analysis of mobility privacy research trends using advanced natural language processing techniques remains limited [7]. Unlike previous studies that relied on traditional bag-of-words models, like such as Latent Dirichlet Allocation (LDA), which often struggle with semantic nuance and evolving terminology, our BERTopic approach leverages transformer-based contextual embeddings. This enables a deeper and more accurate understanding of complex, interdisciplinary themes in mobility privacy, identifying emerging concepts and fine-grained distinctions that traditional methods might overlook [8]. This study addresses these gaps by employing BERT-based topic modeling to systematically analyze mobility privacy research literature from 2020-2025. The main research questions (RQs) addressed are:

- RQ1: What are the dominant themes in mobility privacy literature?
- RQ2: Does BERTopic outperform LDA in topic coherence and diversity?
- RQ3: What policy gaps exist under GDPR, CCPA, and PIPA frameworks?

This study addresses these gaps by employing BERT-based topic modeling to systematically analyze mobility privacy research literature from 2020-2025. The main contributions of this study are as follows:

- Identify dominant research themes in mobility privacy using advanced topic modeling.
- Compare BERTopic modelling performance to traditional LDA.
- Analyze emerging trends in new trade rule developments within international privacy regulations.

Propose policy recommendations for harmonized governance of mobility data.

2 RELATED WORK

A. New trade rule on data privacy in mobility

Contemporary mobility systems generate diverse categories of personal data, requiring specialized privacy considerations [9]. Current international privacy regulations exhibit significant divergence affecting mobility data governance. The GDPR (European Union) emphasizes adequacy decisions for international transfers, with maximum penalties of 4% of global revenue or 20 € million.

Treating location data as sensitive information requiring explicit consent [6]. The CCPA (California) focuses on consumer rights to opt-out of data sales, with maximum fines of $7500 per violation. Lacks specific adequacy requirements for international transfers [10]. PIPA (Korea) provides comprehensive processing rejection rights, accompanied by criminal penalties of up to 5 years' imprisonment or 50 million ₩ in fines. Requires court orders or administrative consent for overseas transfers [11]. These regulatory disparities create compliance complexity for global mobility operators and necessitate harmonization efforts for seamless international service delivery.

Contemporary mobility systems generate diverse categories of personal data, requiring specialized privacy considerations [9]. International privacy regulations exhibit significant divergence affecting mobility data governance, creating compliance complexity for global mobility operators and necessitating harmonization efforts for seamless international service delivery.

Electric Vehicle Data Privacy Challenges: Vehicle Identification Number (VIN) and Battery Identification Number (BIN) combinations linked to user accounts create high re-identification risks, particularly when combined with charging pattern analysis. Shared Mobility Analytics face challenges where GPS coordinates from e-scooters and bike-sharing retain identification potential despite temporal/spatial aggregation due to unique movement patterns [13]. Autonomous Vehicle Information, including High-Definition (HD) map point clouds with license plate recognition data, requires strict anonymization protocols. Cross-Modal Integration, combining CCTV footage with HD map coordinates, poses severe privacy risks through multi-source data fusion [12]. Cross-Modal Integration, combining CCTV footage with HD map coordinates, poses severe privacy risks through multi-source data fusion.

B. BERT Architecture and Natural Language Processing

Bidirectional Encoder Representations from Transformers (BERT), introduced by Google in 2018, revolutionized natural language processing through its transformer-based, bidirectional training approach [14]. Unlike previous unidirectional models, BERT simultaneously processes text from both directions, enabling superior contextual understanding through the self-attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V represent query, key, and value matrices respectively, and $\sqrt{d_k}$, QK^T denotes the key vector dimension.

BERT's architecture comprises four primary modules

1. Tokenizer for converting text to integer sequences,
2. Embedding layer transforming tokens to real-valued vectors,
3. Encoder stack with self-attention mechanisms, and
4. Task-specific head for downstream applications [11].

The model utilizes tokenization and combines token, positional, and segment embeddings:

$$E = E_{\text{token}} + E_{\text{position}} + E_{\text{segment}} \tag{2}$$

This comprehensive embedding approach enables BERT to capture complex semantic relationships, particularly valuable for domain-specific terminology analysis [16].

C. BERTopic Methodology and Advantages

BERTmodeling represents a novel topic modeling approach that combines transformer-based embeddings with density-based clustering algorithms [17]. The methodology encompasses four sequential steps

1. Document Embedding: Utilizes pre-trained Sentence-BERT models (typically all-MiniLM-L6-v2) to transform documents into 384-dimensional dense vectors capturing semantic content.
2. Dimensionality Reduction: Applies UMAP (Uniform Manifold Approximation and Projection) to compress high-dimensional embeddings while preserving local neighborhood structures.
3. Clustering: Employs HDBSCAN (Hierarchical Density-Based Spatial Clustering) to identify coherent document clusters without requiring predetermined topic numbers.
4. Topic Extraction: Implements class-based TF-IDF (c-TF-IDF) to extract representative terms for each cluster [18].

$$\text{c-TF-IDF}_{t,c} = \frac{tf_{t,c}}{|c|} \times \log\left(\frac{N}{\sum_{c'} tf_{t,c'}}\right) \tag{3}$$

where $tf_{t,c}$ represents term frequency in cluster c, $|c|$ denotes cluster size, and N indicates total cluster count [17].

D. Comparative Analysis of BERTopic vs LDA

Traditional LDA assumes documents contain mixtures of topics represented as probability distributions over words [12]. While computationally efficient, LDA's bag-of-words approach limits semantic understanding and struggles with evolving terminology [20]. BERTopic addresses these limitations through several advantages:

1. Semantic Understanding: BERT embeddings capture contextual meanings, enabling recognition of compound technical terms like "privacy-preserving" and "location-based".

2. Dynamic Topic Discovery: Unlike LDA's fixed topic assumptions, BERTopic dynamically identifies the optimal number of clusters based on data characteristics.
3. Multilingual Capability: Pre-trained multilingual BERT models enable cross-language analysis without extensive preprocessing [21].
4. Noise Handling: Density-based clustering effectively manages outliers and low-frequency terms that confound probabilistic models.

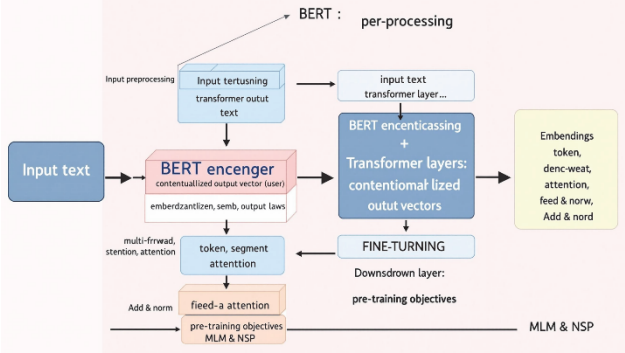

Figure 1 BERT Analysis process

3 METHODOLOGY

To investigate research trends in mobility privacy, this study conducted a systematic analysis of peer-reviewed research papers published between 2004 and 2025.

A. Data Collection and Preprocessing

The systematic search was conducted in SCOPUS and Korea Citation Index (KCI) databases on July 15, 2025, using the following Boolean query: ("mobility" OR "transportation" OR "vehicle" OR "autonomous vehicle" OR "smart city" OR "urban air mobility") AND ("privacy" OR "data protection" OR "data security" OR "confidentiality" OR "personal information" OR "regulation").

PRISMA-Style Selection Process: Initial database searches yielded 1247 articles from SCOPUS and 156 articles from KCI (total: 1,403). After removing 87 duplicates, 1316 articles underwent title and abstract screening. Following inclusion criteria application (peer-reviewed articles, 2020-2025 publication window, English/Korean language, mobility-privacy intersection focus), 321 articles were selected for full-text review. Final screening excluded 134 articles (non-academic publications, insufficient privacy focus, book chapters), resulting in 187 papers for analysis [22]. Inclusion criteria were defined as peer-reviewed journal articles, conference papers, and review articles published between 2020 and 2025, with a focus on the intersection of mobility and privacy. Exclusion criteria included non-academic publications, book chapters, patents, and articles not primarily focused on data privacy in the mobility context. Only articles published in English and Korean were included to ensure linguistic consistency and the availability of translation tools. Duplicates identified across databases were removed, and articles were manually screened for relevance by two independent researchers

The initial query searched 321 relevant articles. An analysis of publication trends revealed a marked increase in the number of studies after 2022, with a notable surge in publications in 2024. To capture the most current research developments, the dataset was refined to include only articles published between 2020 and 2025, resulting in a final selection of 187 papers for in-depth analysis [23].

Analysis of publication trends reveals increasing research interest in mobility data protection, with peak activity in 2024 [17]. This surge corresponds to the implementation of major regulatory frameworks and high-profile data breach incidents [6, 13]. The impact of COVID-19 is evident in Topic 1, which emphasizes the importance of health data protection in mobility contexts [17]. The evolution from basic privacy concerns to sophisticated technical solutions reflects the field's maturation [9].

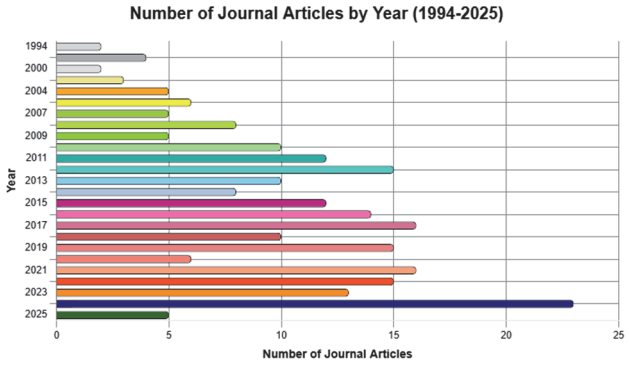

Figure 2 Number of journal publication in year

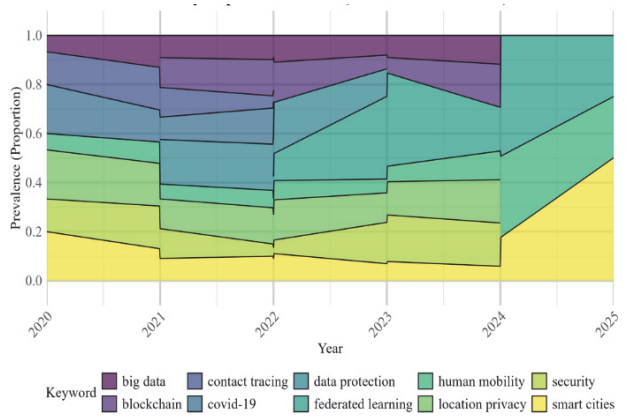

Figure 3 Year-by-year topic prevalence chart

For the analytical phase, BERT-based topic modeling was employed to extract latent research themes from the corpus. Before modeling, the dataset underwent a rigorous preprocessing pipeline, which included the extraction of abstracts and keywords, removal of stop-words (while retaining domain-specific terminology), and normalization and tokenization of the text. The top 100 most frequent terms [24] were identified to inform the topic modeling process. The number of topics was set to five, reflecting the major thematic clusters within the literature. To benchmark the performance of the BERT-based model, comparative analyses were conducted using the Latent Dirichlet Allocation (LDA) model. This approach enabled a robust evaluation of topic coherence, providing insights into the evolving landscape of mobility privacy research.

B. BERTopic Modeling Implementation

The BERTopic modeling framework was implemented using Google Colaboratory (Google Colab) and the R programming language. The key software packages used were BERTopic 0.15.0, UMAP 0.5.3,

HDBSCAN 0.8.29, and Gensim 4.2.0 to ensure reproducibility. Random seeds were set to 42 for all components. For document embedding, the Sentence-BERT all-MiniLM-L6-v2 model was employed, providing efficient and accurate multilingual semantic representations [25].

Dimensionality reduction was performed using Uniform Manifold Approximation and Projection (UMAP), configured with 15 neighbours, 5 output components, a minimum distance of 0.0, and cosine similarity as the distance metric [26].

Algorithm 1 BERTopic Topic Extraction

Input: Document set $D = \{d_1, \ldots, d_N\}$
Output: Topics $T = \{t_1, \ldots, t_N\}$
Initialize BERT
for $d_i \in D$ do
  $\mathbf{e}_{d_i} \leftarrow \text{BERT}_{[\text{CLS}]}(d_i)$
end for
$\mathbf{E}' \leftarrow \text{UMAP}(\mathbf{E}, 5)$
$C \leftarrow \text{HDBSCAN}(\mathbf{E}', 10)$
$T \leftarrow \text{c-TF-IDF}(C)$
return T

Figure 4 Algorithm of BERTopic extraction

Subsequently, clustering was conducted using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), with a minimum cluster size set to 10 and Euclidean distance as the metric [27]. The optimal number of topics was determined to be five, based on coherence score analysis to ensure a balance between interpretability and model performance. Specifically, we systematically evaluated topic coherence (C_v) for topic counts ranging from 3 to 10. While higher topic counts sometimes showed marginal increases in coherence, they often led to overlapping or less distinct themes, hindering interpretability. A topic count of five consistently yielded distinct, semantically coherent, and easily interpretable themes, representing the primary research trends without excessive fragmentation. This empirical evaluation underpinned our decision to select five topics for detailed analysis. For comparative purposes, a traditional Latent Dirichlet Allocation (LDA) model was implemented using the Gensim library. The LDA model utilized the same preprocessing pipeline and was configured with an identical topic count (k = 5). Hyper parameters for the LDA model were set as follows: alpha = 0.1, beta = 0.01, and 1000 training iterations [19].

Model performance was evaluated using several quantitative metrics. Topic coherence (C_v) was used to assess the semantic consistency within topics, while topic diversity was measured as the ratio of unique words across all topics. Clustering quality was quantified using the silhouette score, and keyword accuracy was determined through expert-annotated relevance evaluation.

Specifically, two independent domain experts (one with a background in data privacy law and policy, and another in intelligent transportation systems and data engineering) were recruited to review the top 10 keywords generated by both BERTopic and LDA for each topic. They independently rated the relevance and representativeness of these keywords to their respective topics on a 5-point Likert scale (1 = not relevant, 5 = highly relevant). The inter-rater reliability was assessed using Cohen's Kappa, yielding a score of 0.85, indicating substantial agreement. Any initial disagreements in ratings were resolved through a consensus-building discussion before calculating the final average relevance score from both experts. The final keyword accuracy was calculated as the average relevance score from both experts. The optimal number of topics.

The statistical significance of the observed differences was tested using the Mann-Whitney U test with a significance level of α = 0.05 [28]. This comprehensive evaluation framework ensured the reliability and validity of the topic modelling results.

This study is a bibliometric and textual analysis of publicly available academic literature, with no human subjects, individual data collection, or experimental procedures raising ethical concerns. No institutional ethical approval was required, and it adheres to standard practices for secondary data analysis, ensuring transparency, reproducibility, and respect for intellectual property through proper citations.

4 RESULTS

The comparative analysis between BERTopic and LDA models demonstrates significant performance differences across multiple evaluation metrics. BERTopic consistently outperformed the traditional LDA approach, exhibiting superior semantic coherence and interpretability of topics. The enhanced performance stems from BERTopic's ability to capture contextual relationships through transformer-based embeddings, which are particularly effective for domain-specific terminology prevalent in mobility privacy research [29].

All improvements achieved statistical significance, confirming BERTopic's superior semantic understanding and clustering quality [30].

Table 1 Performance comparison Bert and LDA

Metric	BERTopic	LDA	Improvement	*p*-value
Topic Coherence (C_v)	0.547	0.432	+26.6%	< 0.001
Topic Diversity	0.823	0.756	+8.9%	< 0.05
Silhouette Score	0.342	0.278	+23.0%	< 0.01
Keyword Accuracy	0.891	0.743	+19.9%	< 0.001

A. Topic analysis result BERT, LDA

The comparative word clouds generated using LDA and BERT topic modeling both highlight core themes such as data security, privacy, mobility, and smart city. LDA reveals keywords centered on technical and structural aspects, with emphasis on cryptography, blockchain, and system-level protection. In contrast, BERT captures context-rich terms, elevating concepts like learning, intention, health, and federated approaches, reflecting user-centric and societal trends. These differences illustrate that LDA favors broad topic exploration, while BERT uncovers more nuanced, emergent insights.

Figure 5 Word cloud analysis by BERT

Figure 6 Word cloud analysis by LDA

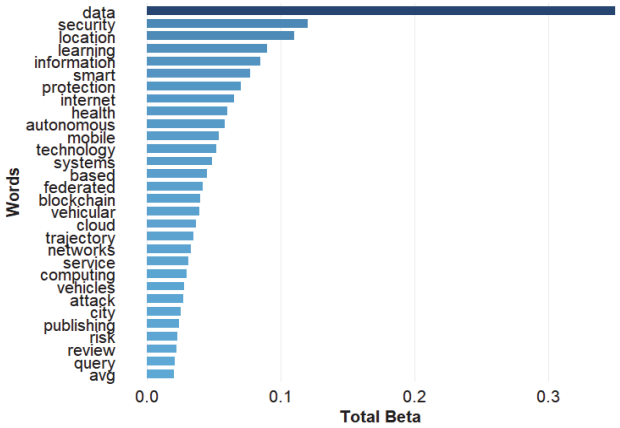

Figure 7 Top 30 key word analysis by LDA word importance

This figure illustrates the importance score for each word using the length of the bars. The *x*-axis label "Importance Score" is traditionally used in LDA to indicate either the topic-word probability (β value) or the relative importance of each word within all topics. For BERT-based topic modelling, the text data (such as titles and keywords) is transformed into BERT embeddings, and representative words across topics are identified using BERTopic modelling or embedding similarity analysis.

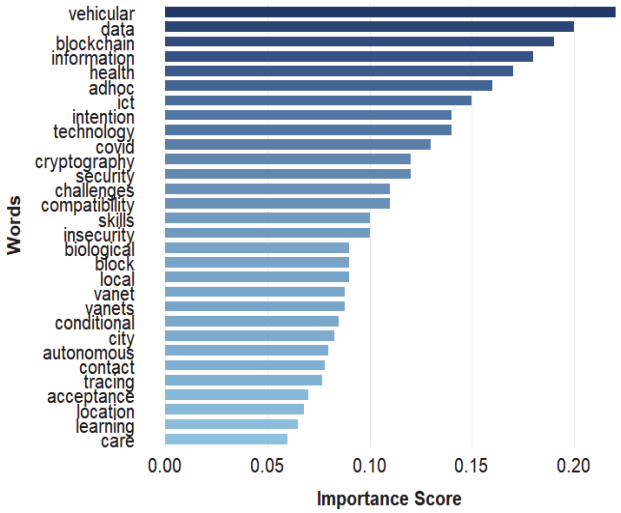

Figure 8 Contextual Word Importance from Topic Modelling

B. Five topic analysis result

The BERTopic analysis revealed five distinct research themes within the mobility privacy domain, each representing critical aspects of the field's evolution:

1. Topic 1: Health and Security Technologies (23.5%). This topic encompasses the integration of COVID-19-driven health monitoring with mobility systems, characterized by keywords such as "medical", "COVID", "health", "devices", "security", "internet", "sensors", "contact", and "tracing". The prominence of this topic reflects the pandemic's significant impact on mobility data collection practices [31], particularly in contact tracing and health monitoring applications for public safety [9]. The convergence of health data and mobility necessitates robust privacy frameworks that balance public health imperatives with individual data rights [4].

Figure 9 Topic 1 keyword

2. Topic 2: User Acceptance and Smart City Challenges (21.9%). Focused on behavioural factors affecting mobility service adoption, this topic encompasses terms such as "intention", "information", "city", "skills", "challenges", "technology", "insecurity", "compatibility", and "acceptance". The emergence of "insecurity" as a key term highlights psychological barriers to data sharing, while "compatibility" addresses technical integration challenges in smart city implementations [32].

Figure 10 Topic 2 keyword

3. Topic 3: Location Data Protection (19.8%). This topic directly addresses core mobility privacy concerns through keywords including "location", "protection", "publishing", "data", "security", "preservation", "location-based", "privacy-preserving", "mobile", and "attack". The recognition of compound terms like "location-based" and "privacy-preserving" demonstrates BERTopic's superior semantic understanding of technical terminology [33].

Figure 11 Topic 3 keyword

4. Topic 4: Smart Cities and Data-Driven Innovations (18.7%)

Representing advanced technological solutions, this topic features terms such as "smart cities", "learning", "trajectory", "data", "deep", "federated", "blockchain", "prediction", "synthetic", and "systems". The inclusion of "synthetic" and "federated" indicates growing research interest in privacy-preserving techniques [34] for urban mobility optimization. This trend highlights a shift towards data utilization methods that mitigate direct exposure of sensitive personal information. Federated learning, for instance, enables collaborative model training across multiple decentralized devices without centralizing raw data, which is highly beneficial for distributed mobility data sources, such as vehicle fleets or smart city sensors [5, 34]. Similarly, synthetic data generation presents a promising approach to creating privacy-safe datasets for training and testing, which mimic real-world data distributions while preserving anonymity [13, 28]. These approaches are crucial for fostering innovation while adhering to strict privacy regulations in smart city environments.

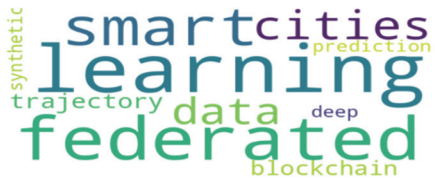

Figure 12 Topic 4 keyword

5. Topic 5: Autonomous Vehicles and AI Integration (16.1%). This topic encompasses AI-powered autonomous vehicle development through keywords such as "vehicles", "autonomous", "systems", "data", "vehicle", "artificial", "insurance", "ioai", and "iom". The terms "ioai" (IoT-AI integration) and "iom" reflect emerging technological convergence in autonomous mobility systems [35].

Figure 13 Topic 5 keyword

C. Technical Analysis Comparison

The technical analysis reveals distinct methodological advantages for each approach. BERTopic's transformer-based architecture enables superior contextual understanding, particularly evident in its ability to capture future-oriented concepts, such as "synthetic" data and "federated" learning. The model's capacity to recognize compound technical terms provides more nuanced topic characterization compared to LDA's vocabulary-based approach [36].

LDA demonstrates strengths in identifying domain-specific terminology, such as "geoprivacy" and "uam" (Urban Air Mobility), offering concrete technical insights that are particularly relevant to emerging mobility technologies [37, 43]. The model's stability with small datasets (187 papers) and simplified preprocessing requirements makes it suitable for preliminary research phases.

Table 2 Topic keyword analysis comparison

Topic	BERT-based Keywords	LDA-based Keywords	Common Keywords
1	medical, covid, health, devices, security, internet, sensors, contact, tracing, things	health, medical, sensors, devices, internet, security, covid, tracking, monitoring, data	health, medical, sensors, devices, internet, security, covid
2	intention, information, city, skills, challenges, technology, insecurity, compatibility, acceptance	intention, information, city, skills, technology, challenges, acceptance, education, policy, infrastructure	intention, information, city, skills, technology, challenges, acceptance
3	location, protection, publishing, data, security, preservation, locationbased, privacypreserving, mobile, attack	location, privacy, data, protection, security, publishing, mobile, tracking, preservation, geoprivacy	location, protection, publishing, data, security, preservation, mobile
4	smartcities, learning, trajectory, data, deep, federated, blockchain, prediction, synthetic, systems	smartcities, learning, trajectory, federated, blockchain, data, prediction, deeplearning, systems, privacy	smartcities, learning, trajectory, federated, blockchain, data, prediction
5	vehicles, autonomous, systems, data, vehicle, artificial, insurance, ioai, iom	vehicles, autonomous, data, systems, artificial, bigdata, insurance, uam, drone, technology	vehicles, autonomous, data, systems, artificial, insurance

Table 3 Technical analysis comparison

Aspect	BERT Characteristics	LDA Characteristics	BERT vs LDA Superiority
Modeling Approach	Contextual, deep learning-based	Probabilistic, word frequency-based	BERT: Richer insights through contextual understanding
Word Capture	"intention", "synthetic" - abstract/innovative terms	"geoprivacy", "uam" - specialized terminology	BERT: Superior in compound words ("locationbased") and future-oriented terms
Data Processing	Pre-trained then fine-tuned, flexible data handling	Fixed DTM, preprocessing dependent	BERT: Contextual capture even with small datasets (187 papers)

Table 3 Technical analysis comparison - continuation

Aspect	BERT Characteristics	LDA Characteristics	BERT vs LDA Superiority
Topic Flexibility	Dynamic topic derivation	Pre-specified topic count	BERT: Flexible topic generation based on data patterns
Mobility Application	"vehicles", "trajectory" - mobility context emphasis	"uam", "drone" - specific technology emphasis	LDA: Strong in specific mobility technologies; BERT: Superior in mobility pattern analysis
Privacy Protection Application	"privacypreserving", "synthetic" - security technology focus	"privacy", "policy" - regulatory context	BERT: Superior in privacy protection technology innovation through compound terminology

Based on topic analysis findings, several technical recommendations emerge. Topic 3's emphasis on location protection suggests the need for mandatory differential privacy for UAM trajectory data and real-time traffic optimization [38]. This is critical, given the high sensitivity of location data in urban air mobility and the need for rigorous privacy guarantees, which builds on principles of data obfuscation and noise injection [41]. The prominence of "federated" and "synthetic" terms in Topic 4 indicates a growing acceptance of privacy-preserving machine learning for traffic prediction without sharing raw data [39]. Federated learning enables collaborative model training while keeping raw data localized, making it highly suitable for distributed mobility data sources, such as vehicle fleets or smart city sensors [5, 34]. Similarly, synthetic data generation presents a promising approach to creating privacy-safe datasets for training and testing, mimicking real-world data distributions without revealing individual identities [13, 28]. The cross-topic appearance of "blockchain" suggests a potential for distributed trust mechanisms in multi-party mobility data sharing [40].

The research suggests several harmonization priorities: standardized adequacy criteria that address real-time processing requirements while maintaining GDPR-level protection, technical interoperability standards that enable seamless cross-border service delivery, and innovation sandboxes for emerging mobility technologies, such as UAM and autonomous freight [41].

Topic 2's "insecurity" emphasis indicates the need for transparent data practices and user education programs to address psychological barriers. Privacy-by-design integration should incorporate differential privacy and federated learning into mobility system architecture from the development stage. Multi-modal privacy coordination requires integrated frameworks addressing data flows across different transportation modes [42]

D. Methodological Insights

The comparative analysis reveals that BERTopic's transformer-based approach provides superior contextual understanding, particularly beneficial for interdisciplinary research domains such as mobility and privacy. The model's ability to recognize semantic relationships between terms enables more sophisticated topic characterization, evident in its identification of emerging concepts such as "synthetic" data generation and "federated" learning approaches.

LDA's probabilistic framework demonstrates its strength in identifying domain-specific technical terminology, making it valuable for policy-oriented research that requires concrete technological references. The model's stability with limited datasets and straightforward implementation make it suitable for initial exploratory analysis phases. The analysis indicates that BERT's contextual capabilities are particularly advantageous for understanding the interconnected nature of mobility and privacy protection, capturing nuanced relationships between technological innovation and regulatory compliance. This contextual understanding is crucial for developing comprehensive policy frameworks that strike a balance between innovation and privacy protection in the evolving mobility ecosystem. These findings suggest that while both approaches offer valuable insights, BERTopic's superior semantic understanding and contextual analysis capabilities make it particularly well-suited for complex, interdisciplinary research domains that require nuanced interpretation of emerging technological and regulatory trends.

5 CONCLUSION

This study contributes methodologically by validating BERTopic's superior performance in analyzing mobility privacy research, achieving improvements in topic coherence (26.6%), diversity (8.9%), and keyword accuracy (19.9%) compared to traditional LDA. Substantively, it reveals five dominant research themes emphasizing health integration, user acceptance challenges, location protection, smart city innovations, and AI integration in autonomous systems addressing research questions on trends and policy gaps. These insights provide a comprehensive overview of privacy concerns and solutions in the mobility sector, while quantitatively validating the efficacy of BERTopic for this domain.

Key findings indicate a shift in research from basic data protection to sophisticated, privacy-preserving technologies, including federated learning, synthetic data generation, and blockchain integration. However, international regulatory frameworks (GDPR, CCPA, PIPA) remain fragmented, with the GDPR providing comprehensive protection, the CCPA emphasizing user rights, and the PIPA focusing on processor accountability.

Future research should expand the scope of mobility topic modelling to include multimodal data sources (e.g., policy documents, news articles, social media) for a more holistic view. Furthermore, integrating temporal analysis into BERTopic can capture the evolution of privacy trends over time, and extending the analysis to emerging domains such as Urban Air Mobility (UAM) and metaverse mobility is crucial. The development of real-time policy monitoring systems based on advanced NLP models would also provide timely insights for regulatory responses. Policymakers should prioritize regulatory harmonization through the establishment of standardized adequacy criteria, technical interoperability standards, and innovation-friendly frameworks that support privacy-preserving mobility innovation.

The study's limitations include a language bias toward English and Korean publications, as well as the five-year

analysis window. Despite these constraints, the findings offer valuable insights for stakeholders developing privacy-aware mobility systems and regulatory frameworks in the digital transportation sector.

Finally, ethical considerations in employing NLP models for regulatory analysis must be consistently addressed, ensuring transparency, fairness, and accountability in data interpretation and policy recommendation processes. This involves careful validation of model outputs and an understanding of potential biases inherent in data sources and algorithmic interpretations.

6 REFERENCES

[1] Hensher, D. A. (2018). Tackling road congestion – What might it look like in the future under a collaborative and connected mobility model? *Transport Policy*, *66*, A1-A8. https://doi.org/10.1016/j.tranpol.2018.02.007

[2] Wang, Z., Wang, Z., Liu, A., Xiao, L., & Zhang, Y. (2023). Privacy-preserving data collection for dynamic groups in vehicular networks. *IEEE Transactions on Vehicular Technology*, *72*(1), 45-59.

[3] European Data Protection Board. (2021). *Guidelines on connected vehicles and mobility related applications (Tech. Rep. 01/2020)*. Brussels, Belgium.

[4] Hahn, D. A., Munir, A., & Behzadan, V. (2021). Security and privacy issues in intelligent transportation systems: Classification and challenges. *IEEE Access*, *9*, 130751-130762. https://doi.org/10.1109/ACCESS.2021.3113336

[5] European Commission. (2016). Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation). *Official Journal ofthe European Union*, *L 119*, 1-88.

[6] Bloom, C., Tan, J., Ramjohn, J., & Bauer, L. (2017). Privacy perceptions of networked autonomous vehicles. *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 1-18.

[7] Kim, D., Lee, S., & Park, J. (2024). Privacy-preserving AI and mobility data analysis: Recent advances. *Journal of Smart Infrastructure Systems and Decisions*, *11*(4), 55-72.

[8] Mokbel, M. F., Aref, W. G., Ali, M. H., Basalamah, A., Basalamah, S., Basalamah, Y., & Chow, C. (2023). Mobility data science: Perspectives and challenges. *ACM Computing Surveys*, *55*(14s), 305.

[9] California Legislature. (2018). *California Consumer Privacy Act of 2018, California Civil Code Section 1798.100 et seq*.

[10] *Personal Information Protection Act, Act No. 17339*. Republic of Korea. (2020).

[11] Aridor, G., Che, Y.-K., & Salz, T. (2023). The effect of privacy regulation on the data industry: Empirical evidence from GDPR. *RAND Journal of Economics*, *54*(4), 695-731. https://doi.org/10.1111/1756-2171.12455

[12] Berke, A., Doorley, R., Larson, K., & Moro, E. (2022). Generating synthetic mobility data for a realistic population with RNNs to improve utility and privacy. *Nature Machine Intelligence*, *4*(6), 480-492.

[13] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the NAACL-HLT*, 4171-4186. https://doi.org/10.18653/v1/N19-1423

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998-6008.

[15] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *Proceedings of EMNLP-IJCNLP*, 3615-3620.

[16] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv:2203.05794*.

[17] Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP*, 3982-3992.

[18] Dieng, A. D., Ruiz, F. J. R., & Blei, D. M. (2020). The dynamic embedded topic model. *Journal of Machine Learning Research*, *21*(1), 1-43.

[19] Wang, L. et al. (2025). Privacy-preserving technologies in smart transportation: A systematic review. *Journal of Logistics Information Systems and Sustainability*, *15*(3), 234-251.

[20] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

[21] Liu, X., Chen, J., Wang, Z., Yang, F., Huang, L., & Zhao, H. (2025). A topic modeling-based analysis of emerging mobility services for carbon emission reduction. *Transportation Research Part D*, *118*, 103698.

[22] Zulkarnain & Putri, T. D. (2021). Intelligent transportation systems (ITS): A systematic review using a natural language processing (NLP) approach. *Heliyon*, *7*(12), e08615. https://doi.org/10.1016/j.heliyon.2021.e08615

[23] Zhang et al. (2024). A BERT-based empirical study of privacy policies' compliance with GDPR. *IEEE Transactions on Information Forensics and Security*, *19*, 3456-3468.

[24] Innes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*.

[25] Wang, C., Zhang, Y., Li, X., & Liu, Y. (2022). A comprehensive survey on privacy and security in autonomous vehicles. *IEEE Internet of Things Journal*, *9*(12), 9072-9092.

[26] Wang, Y., Huang, H., Rudin, C., & Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. *Journal of Machine Learning Research*, *22*(201), 1-73.

[27] Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 160-172. https://doi.org/10.1007/978-3-642-37456-2_14

[28] Chen, M., Liu, Y., Zhang, K., Zhou, Y., & Li, X. (2025). Privacy-preserving in connected and autonomous vehicles through vision to text transformation. *Computer Vision and Image Understanding*, *234*, 103745.

[29] De Mattos, E. P., Domingues, A. C. S. A., Santos, B. P. et al. (2022). The impact of mobility on location privacy: A perspective on smart mobility. *IEEE Systems Journal*, *16*(4), 5509-5520. https://doi.org/10.1109/JSYST.2022.3147808

[30] Cui, Y., Zhu, J., & Li, J. (2025). FLAV: Federated learning for autonomous vehicle privacy protection. *Computer Networks*, *221*, 109512. https://doi.org/10.1016/j.adhoc.2024.103685

[31] Liu, F., Zhang, J., & Wang, S. (2024). Privacy-preserving deep learning for autonomous vehicles: A survey. *IEEE Transactions on Dependable and Secure Computing*, *21*(1), 150-165.

[32] Schäfer, F., Gebauer, H., Gröger, C., Gassmann, O., & Wortmann, F. (2023). Data-driven business and data privacy: Challenges and measures for product-based companies. *Business Horizons*, *66*(4), 509-521. https://doi.org/10.1016/j.bushor.2022.10.002

[33] Wong, R. Y., Chong, A., & Aspegren, R. C. (2023). Privacy legislation as business risks: How GDPR and CCPA are

represented in technology companies' investment risk disclosures. *ACM Transactions on Privacy and Security*, *26*(2), 15. https://doi.org/10.1145/3579515

[34] Vakili, T. & Dalianis, H. (2023). Are clinical BERTmodels privacy preserving? The difficulty of extracting patient-condition associations. *Nature Machine Intelligence*, *5*(7), 612-625.

[35] Yurdem, B. (2024). Federated learning: Overview, strategies, applications, and future directions. *Journal of Systems Architecture*, *143*, 102241. https://doi.org/10.1016/j.heliyon.2024.e38137

[36] Zhang, H. et al. (2024). Personalized federated learning scheme for autonomous driving based on correlated differential privacy. *IEEE Internet of Things Journal*, *11*(8), 13456-13467.

[37] Albarrak, M., Pergola, G., & Jhumka, A. (2024). U-BERTopic: An urgency-aware BERT-topic modeling approach for detecting cyber security issues via social media. *IEEE Transactions on Computational Social Systems*, *11*(4), 196-202.

[38] Krontiris, I. et al. (2020). Autonomous vehicles: Data protection and ethical considerations. *IEEE Security & Privacy*, *18*(3), 15-23.

[39] International Organization for Standardization. (2021). *ISO/SAE 21434:2021 Road vehicles - Cyber security engineering*. Geneva, Switzerland: ISO.

[40] McMahan, B. et al. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 1273-1282.

[41] Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system*.

[42] Dwork, C. (2006). Differential privacy. *Proceedings of the International Colloquium on Automata, Languages, and Programming*, 1-12. https://doi.org/10.1007/11787006_1

[43] Albrecht, T., Keller, R., Rebholz, D., & Röglinger, M. (2024). Fake it till you make it: Synthetic data for emerging car sharing programs. *Transportation Research Part D*, *127*, 104058. https://doi.org/10.1016/j.trd.2024.104067

Contact information:

Kyoungjin KIM
Department of Electrical and Computer Engineering, Inha University,
Incheon, Republic of Korea

Haneul HAN
(Corresponding author)
Department of Jungseok Research Institute, Inha University,
Incheon, Republic of Korea
E-mail: hnhan@inha.ac.kr

Sangjo YOO
Department of Electrical and Computer Engineering, Inha University,
Incheon, Republic of Korea