Preprints.org

Article

Not peer-reviewed version

Semantic Topic Modeling of Aviation Safety Reports: A Comparative Analysis Using BERTopic and PLSA

Aziida Nanyonga , Keith Joiner , Ugur Turhan , Graham Wild *

Posted Date: 20 May 2025

doi: 10.20944/preprints202505.1509.v1

Keywords: aviation safety; topic modeling; BERTopic; pLSA; ASN reports; text mining

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

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.

Article

Semantic Topic Modeling of Aviation Safety Reports: A Comparative Analysis Using BERTopic and PLSA

Aziida Nanyonga [1], Keith Joiner [2], Ugur Turhan [3] and Graham Wild [3,*]

[1] School of Engineering and Technology, University of New South Wales, Canberra, ACT 2600, Australia
[2] Capability Systems Centre, University of New South Wales, Canberra, ACT 2610, Australia
[3] School of Science, University of New South Wales, Canberra, ACT 2612, Australia
* Correspondence: g.wild@unsw.edu.au

Abstract: Aviation safety analysis increasingly relies on extracting actionable insights from narrative incident reports to support risk identification and improve operational safety. Topic modeling techniques such as Probabilistic Latent Semantic Analysis (pLSA) and BERTopic offer automated methods to uncover latent themes in unstructured safety narratives. This study evaluates the effectiveness of each model in generating coherent, interpretable, and semantically meaningful topics for aviation safety practitioners and researchers. We assess model performance using both quantitative metrics (topic coherence scores) and qualitative evaluations of topic relevance. The findings show that while pLSA provides a solid probabilistic framework, BERTopic leveraging transformer-based embeddings and HDBSCAN clustering produces more nuanced, context-aware topic groupings, albeit with increased computational demands and tuning complexity. These results highlight the respective strengths and trade-offs of traditional versus modern topic modeling approaches in aviation safety analysis. This work advances the application of natural language processing (NLP) in aviation by demonstrating how topic modeling can support risk assessment, inform policy, and enhance safety outcomes.

Keywords: aviation safety; topic modeling; BERTopic; pLSA; ASN reports; text mining

1. Introduction

The analysis of aviation safety reports plays a vital role in identifying recurring hazards, understanding contributory factors, and implementing corrective actions to enhance flight safety [1]. These reports, often prepared by pilots, air traffic controllers, and safety investigators, contain unstructured textual narratives that describe the sequence of events, environmental conditions, and operational decisions leading to aviation incidents and accidents. Such qualitative data, while rich in insight, is challenging to analyze at scale using traditional manual methods. As global aviation activity continues to rise, the accumulation of safety data has become vast and complex, making it necessary to adopt advanced computational methods for processing and interpreting these narratives [2].

In recent years, topic modeling has emerged as a valuable text mining approach for exploring hidden themes and semantic structures within large corpora. By automatically discovering latent topics in textual data, topic modeling allows safety analysts and researchers to group similar terms and uncover prevalent issues across incident reports. This contributes to enhanced situational awareness, data-driven risk assessment, and more informed policy development. Classic topic modeling approaches such as Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (pLSA) have been widely adopted in domains such as healthcare, legal analysis, and social sciences [3–5]. In aviation, these methods have been used to explore patterns in flight safety narratives, pilot reports, and accident databases [6,7].

PLSA, introduced by Hofmann (1999), models the probability of a word given a document through a latent class model that assumes each document is a mixture of topics, and each topic is a

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

distribution over words modeling [8]. While pLSA was foundational in demonstrating the potential of probabilistic models for document clustering, it suffers from several limitations. These include its tendency to overfit, lack of a generative model for new documents, and difficulty scaling to large corpora. Additionally, pLSA's bag-of-words assumption fails to capture semantic relationships and contextual meanings, which are especially important in safety-critical narratives where subtle language nuances can signify distinct operational risks.

To address these shortcomings, modern topic modeling approaches have begun leveraging recent advances in deep learning and language modeling. One such model is BERTopic, which integrates transformer-based embeddings (e.g., BERT) with clustering algorithms such as Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) and dimensionality reduction techniques like Uniform Manifold Approximation and Projection (UMAP). By generating dense vector representations of text that retain contextual information, BERTopic enables dynamic topic extraction with higher semantic accuracy and interpretability [9,10]. This contextual awareness is critical in domains like aviation safety, where reports may include domain-specific jargon, evolving operational conditions, and intricate incident sequences. For instance, BERTopic can distinguish between topics like "engine flameout during takeoff" and "engine malfunction during cruise" due to its ability to capture the context surrounding keywords.

This study addresses the research question: "Can BERTopic outperform pLSA in extracting relevant and interpretable topics from aviation safety reports?" To explore this, a proprietary dataset obtained from the Aviation Safety Network (ASN) is utilized. The dataset comprises detailed narratives of aviation safety occurrences, categorized according to various damage levels. These reports provide an invaluable resource for analyzing and comparing topic modeling techniques across a range of incident scenarios, including mechanical failures, airspace intrusions, and pilot misjudgments. By applying both pLSA and BERTopic, the study aims to evaluate their effectiveness in uncovering latent themes that can enhance understanding of aviation safety issues.

The contributions of this study are threefold: a) A side-by-side comparison of pLSA and BERTopic is conducted, assessing their ability to generate coherent, distinct, and meaningful topics from ASN narratives. B) Both models are evaluated using a combination of quantitative metrics (e.g., coherence scores) and qualitative analysis, including expert validation and domain relevance of the extracted topics. C) The broader implications of topic modeling for aviation safety research are discussed, with emphasis on automation, risk categorization, and the potential integration of such models into Safety Management Systems (SMS).

By providing a comparison of these two models, this research aims to assist data scientists, aviation safety analysts, and policymakers in selecting the most effective method for analyzing narrative-based safety data. Ultimately, the findings contribute to the growing body of literature on natural language processing (NLP) applications in aviation safety and demonstrate the promise of context-aware topic modeling for risk detection and operational improvement.

The remainder of the paper is organized as follows: Section 2 presents a review of relevant studies in topic modeling, particularly in the context of aviation safety. Section 3 details the methodology, including dataset characteristics and evaluation criteria. Section 4 discusses the experimental results, while Section 5 provides a detailed analysis. Finally, Section 6 concludes with insights and future research directions.

2. Related Work

Topic modeling has become a critical technique in Natural Language Processing (NLP) for uncovering latent semantic structures within large collections of textual data. Over the years, various methods have been proposed to perform topic modeling, each offering unique strengths and weaknesses. This section provides an overview of key research in topic modeling, focusing on two influential approaches: pLSA and BERTopic, along with their applications in aviation safety and other domains.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

pLSA is one of the seminal probabilistic approaches to topic modeling, introduced by Hofmann (1999) [8]. pLSA models the co-occurrence of terms and documents through the introduction of a latent variable that represents topics. The underlying assumption of PLSA is that each document is a mixture of topics, where each topic is characterized by a probability distribution over words. This probabilistic framework enables the extraction of hidden thematic structures from large datasets, facilitating the analysis of textual data in diverse domains [11].

Despite its foundational role in topic modeling, pLSA exhibits several limitations. One of the primary drawbacks is its tendency to overfit, especially when applied to smaller datasets, due to its reliance on a fixed generative model for new documents [12,13]. Additionally, PLSA employs the bag-of-words (BoW) model, which simplifies word relationships and fails to capture the semantic context of words crucial issues when dealing with complex domains like aviation safety. This limitation results in topics that may lack meaningful semantic relationships, making pLSA less suitable for tasks where context and domain-specific knowledge are critical.

Despite these challenges, pLSA has been successfully applied in various fields, such as bioinformatics, where it was used to analyze gene sequence data [14,15], and information retrieval [16], where it has been used to extract meaningful themes from large-scale document collections. Its contributions to the development of topic modeling have shaped subsequent advancements in this area.

In contrast to pLSA, BERTopic represents a more recent advancement in topic modeling that incorporates transformer-based embeddings combined with clustering techniques to generate dynamic and context-aware topics. BERTopic leverages sentence embeddings from transformer models like BERT (Bidirectional Encoder Representations from Transformers), which capture semantic relationships between words and phrases [17–19]. This enables BERTopic to generate more coherent, interpretable, and domain-relevant topics compared to traditional methods such as pLSA and Latent Dirichlet Allocation (LDA)[20].

BERTopic's ability to use BERT embeddings allows it to capture contextual relationships that are important for understanding complex and evolving language. This is especially beneficial in domains like aviation safety, where terminology and incident descriptions often involve intricate details, such as "mechanical failure," "pilot misjudgment," or "airspace incursion" [9]. Additionally, BERTopic employs UMAP or HDBSCAN for dimensionality reduction and clustering, which allows it to handle high-dimensional data effectively and uncover meaningful topics even in noisy or sparse datasets [21].

Studies have demonstrated BERTopic's superiority in domains requiring nuanced understanding, such as healthcare, where it has been used to identify themes from clinical notes [22], finance, where it was applied to analyze financial documents [23], and social media analysis, where it has been used to extract insights from large volumes of user-generated content [24]. BERTopic's ability to model the evolving nature of language makes it particularly suitable for complex domains like aviation safety, where the language used in safety reports can change over time as new technologies and practices emerge.

The application of topic modeling in aviation safety is an emerging field. Aviation safety reports, particularly those produced by ASN, contain valuable unstructured data that detail incidents and accidents, providing insights into systemic safety issues. These reports often describe critical events such as mechanical failures, human errors, and operational deficiencies, all of which are essential for improving safety protocols and preventing future occurrences [25].

Earlier studies have employed methods like LDA to extract themes from aviation safety data. LDA, a generative model that assumes documents are mixtures of topics and topics are mixtures of words, has been widely used in various text mining applications. However, its reliance on simplistic text representations often leads to less coherent topics and struggles to capture domain-specific nuances. For example, LDA has been used to model aviation safety reports, but it frequently generates topics that lack clarity and interpretability due to its oversimplified assumptions [26]. Similarly, pLSA has been applied in early studies of aviation safety reports but has faced challenges

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

in scalability and fails to account for the complex relationships within the text, particularly when interpreting industry-specific terminology like "aeronautical hazards" or "pilot deviation" [27].

Recent advancements, including the introduction of BERTopic, have shown considerable promise in aviation safety research. For instance, in 2022, Grootendorst demonstrated the effectiveness of BERTopic in analyzing complex datasets, including safety and risk analysis reports. Compared to traditional methods like PLSA, BERTopic provides better scalability, coherence, and interpretability, making it more suitable for analyzing the detailed and often lengthy narratives found in aviation safety reports [28]. The ability of BERTopic to adapt to the evolving language and context of aviation safety reports makes it an ideal tool for identifying emerging safety issues, monitoring trends, and categorizing risks.

While both pLSA and BERTopic have demonstrated success in different domains, few studies have directly compared these two methods, especially in the context of aviation safety. A comparative study by Ibraimoh et al., [29] highlighted the superiority of BERTopic over pLSA in terms of coherence and interpretability when applied to datasets such as stock overflow incidents. Similarly, research has pointed out that transformer-based models like BERT outperform traditional probabilistic methods when applied to complex textual data, such as customer support documents and incident reports [6,30].

However, there remains a gap in the literature regarding direct comparisons between pLSA and BERTopic in aviation safety contexts. The few available studies that investigate topic modeling in aviation safety focus on traditional models and do not fully explore the advantages of transformer-based techniques like BERTopic in this domain. This study seeks to fill this gap by providing a detailed comparison of pLSA and BERTopic in extracting relevant and coherent topics from aviation safety reports, particularly focusing on the ASN dataset.

Building upon previous studies, this research aims to compare the effectiveness of pLSA and BERTopic in extracting meaningful and coherent topics from aviation safety reports. The study emphasizes the relevance, coherence, and domain-specific applicability of the models when applied to the ASN dataset, which contains detailed and categorized reports on aviation incidents. This comparison is expected to offer new insights into the strengths and limitations of both approaches in the context of aviation safety data, providing valuable contributions to the growing body of literature on NLP and topic modeling in aviation research.

3. Materials and Methods

This section provides a comprehensive outline of the methodology employed to evaluate and compare the performance of pLSA and BERTopic in extracting meaningful topics from aviation safety reports. The process consists of data collection and preprocessing, followed by the implementation of the two topic modeling techniques. The evaluation of their effectiveness is conducted using several performance metrics, as depicted in Figure 1.

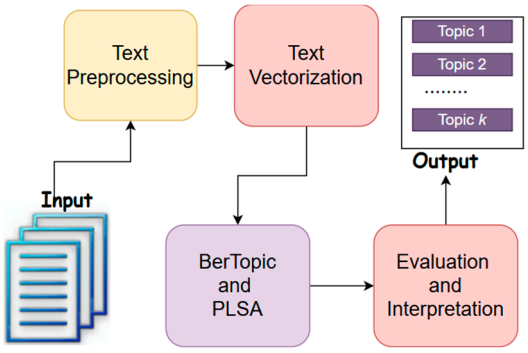

Figure 1. Methodological framework.

A. Data Collection

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

Aviation incident and accident investigation reports are rich sources of information that detail the nature and causes of aviation safety events. Various organizations publish these reports, including ASN, Australian Transport Safety Bureau (ATSB), the Aviation Safety Reporting System (ASRS), and the National Transportation Safety Board (NTSB). For this study, the focus was specifically placed on the ASN aviation incident and accident investigation reports. The dataset utilized in this study covers reports from 2013 to 2022, spanning a decade of aviation safety events. The data was directly sourced from the ASN website, and the dataset consists of 4875 records. The dataset includes detailed narratives of incidents and accidents, categorized by damage levels (e.g. Damaged beyond repair, Missing, Substantial, Destroyed, None, Minor, Unknown). After data cleaning and preprocessing steps, a refined dataset of 4282 records was created. This dataset, which serves as the core data for the analysis, includes the 'Narrative and 'Damage Level' fields, which capture the textual content of the reports and the severity of damage level of the aircraft. These two fields provide rich, unstructured data that is ideal for topic modeling techniques such as pLSA and BERTopic.

B. Data Processing

Before applying topic modeling techniques, the raw text data required preprocessing to ensure consistency and quality for analysis. The preprocessing steps involved several key stages to clean and transform the data into a suitable format for topic modeling: the first step was tokenization, where the text was split into individual words or tokens using the Natural Language Toolkit (NLTK). Tokenization is a crucial step NLP that prepares the text for further analysis by breaking it down into manageable units. All text was converted to lowercase to standardize the tokens and remove any case sensitivity that could affect the consistency of the topic modeling process. Commonly used words such as "the," "and", "or", "of," etc., which carry little semantic value, were removed using the NLTK stopword list. Removing these stopwords ensures that the model focuses on the more meaningful terms in the text. The next step was lemmatization, where the text was lemmatized using the WordNetLemmatizer from NLTK, which reduced words to their base or root forms. For example, words like "running" and "ran" were reduced to the base form "run," thus improving consistency in the text and ensuring that variations of the same word are treated as a single token. Finally, all special characters, punctuation marks, numbers, and other non-alphabetic symbols were filtered out to ensure that the dataset only contained meaningful textual content. After these preprocessing steps, the text data was transformed into a clean and uniform corpus. This ensured that both pLSA and BERTopic received identical input, making it possible to conduct a fair and valid comparison between the two topic modeling approaches.

C. Topic Modeling Procedure

Once the data was preprocessed, it was ready for topic modeling. Both pLSA and BERTopic were applied independently to extract latent topics from the dataset, as outlined in the methodology framework (Figure 1). For PLSA, the process began with the creation of a document-term matrix (DTM), which is a sparse matrix where each row represents a document, and each column represents a unique word in the corpus. The values in the matrix represent the frequency of terms in each document. Model fitting was carried out using the Expectation-Maximization (EM) algorithm, which iteratively estimates the topic-word and document-topic distributions. These distributions were used to identify the topics in the corpus, each represented as a probabilistic distribution over words. The resulting topics were then evaluated for coherence and relevance. On the other hand, BERTopic uses transformer-based models to generate high-dimensional embeddings. The text was first converted into embeddings using pre-trained models like BERT, which represent the semantic content of the documents in a high-dimensional vector space. These embeddings were then clustered using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm. This clustering method groups the embeddings that are close together in the vector space, typically corresponding to documents that are thematically related. Once the clusters were formed, dynamic

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

topic representation was used to label each cluster with interpretable topics. This process involves extracting the most representative words from each cluster and assigning them as labels that best describe the topic. These two topic modeling techniques, pLSA and BERTopic, were thus implemented independently to extract meaningful topics from the aviation safety reports dataset. Figure 2 illustrates the distribution and relationships among the identified topics. Each circle in the map corresponds to a topic, with its size indicating the frequency of topic occurrence. The D1 and D2 axes represent the two-dimensional space generated by UMAP, which is employed to visualize topic similarity based on their underlying embeddings. Additionally, the map demonstrates that once the initial topics are identified, an automated topic reduction process can be re-applied to refine the topic structure further.

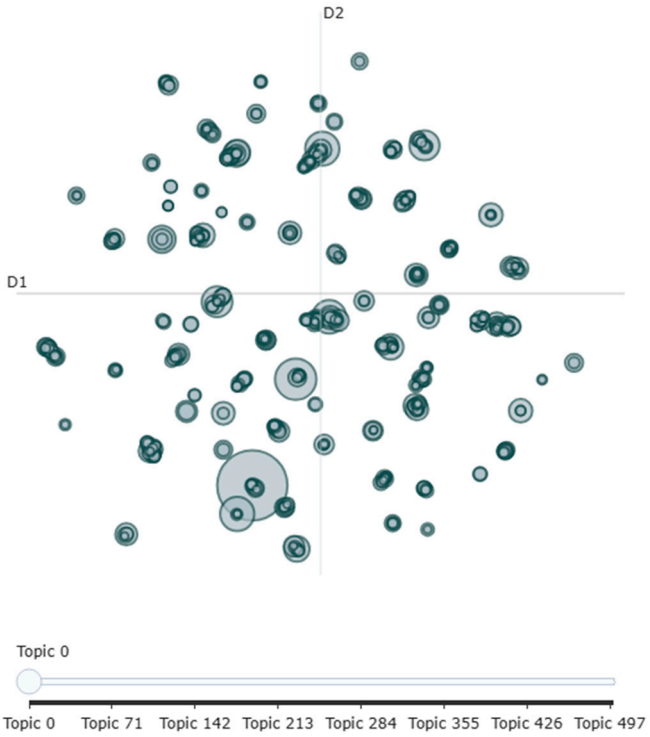

Figure 2. BERTopic's interactive intertopic distance map.

D. Evaluation Metrics

To evaluate and compare the performance of pLSA and BERTopic, several metrics were used. Topic coherence, which measures the semantic similarity between words within a topic, was a key metric in this study. A higher coherence score indicates that the words in the topic are more semantically related, which typically results in more interpretable topics. The C_v metric from Gensim's Coherence Model was used to compute topic coherence, as it has been shown to align well with human judgment of topic quality. In addition to coherence, interpretability was assessed through a manual inspection process by aviation safety experts. These experts reviewed the topics generated by both pLSA and BERTopic and evaluated their clarity and relevance in the context of aviation safety. The experts focused on how well the topics captured meaningful patterns from the data. Finally, scalability was evaluated by measuring the computational efficiency of each model, specifically the time required for training the models and the memory usage during the execution of topic modeling [31]. This provided insights into how well each method handles datasets, such as the 4282 records in this study.

E. Experimental Setup

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

The experiments were conducted in a Python-based environment with a system configuration that included an Intel i7 processor, 32GB RAM, and an NVIDIA GPU to facilitate the efficient processing of transformer-based embeddings (for BERTopic). The software environment used included Python 3.10, along with the following libraries: BERTopic 0.13.0, Gensim 4.3.1, and NLTK 3.8.0. In the case of PLSA, preliminary experiments indicated that setting the number of topics to 6 would be optimal for the dataset. For BERTopic, the minimum cluster size was set to 15, and default transformer embeddings from BERT were utilized. These configurations allowed for a consistent and fair evaluation of both methods across the dataset. The results of these experiments were analyzed to determine which of the two methods pLSA or BERTopic was better suited for extracting meaningful and coherent topics from aviation safety reports, particularly in terms of their relevance, coherence, and domain-specific applicability within the context of the ASN dataset.

4. Results

This section presents the findings from the application of two topic modeling techniques, pLSA and BERTopic, on a curated dataset obtained from the ASN. The analysis aimed to identify latent themes within narrative descriptions of aviation incidents and accidents, categorized by aircraft damage severity.

4.1. Model Performance Metrics

The performance of the pLSA and BERTopic models was assessed using two widely accepted evaluation metrics: topic coherence and perplexity. The pLSA model achieved a coherence score of 0.7634 and a perplexity of −4.6237, indicating more statistically consistent and semantically meaningful topics. In contrast, BERTopic yielded a lower coherence score of 0.531 but a slightly better perplexity of −5.5377, suggesting improved predictive capability. Despite its lower coherence, BERTopic demonstrated greater interpretability, particularly due to its use of transformer-based embeddings and dimensionality reduction techniques such as UMAP, which facilitate clearer visualization and dynamic adjustment of topic granularity.

4.2. Comparative Topic Analysis

4.2.1. Topic Words and Thematic Labels

Table 1 illustrates the top 10 words for each topic derived by both models, along with assigned thematic labels based on manual inspection and cross-validation by aviation domain experts. Their expertise ensured that the extracted topics aligned with real-world operational contexts and safety concerns, improving the reliability of thematic interpretation. While both models extracted topics related to aviation incidents, BERTopic consistently yielded more semantically cohesive and domain-relevant themes (e.g., Bird Strike Investigations, Helicopter Operations, Engine Failures), as confirmed by expert reviewers. In contrast, pLSA occasionally grouped semantically disjointed terms (e.g., “aircraft,” “drugs,” “tug”) under a single topic, which complicated thematic labeling. Figures 3 and 4 further illustrates the top terms per topic for BERTopic and pLSA respectively.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

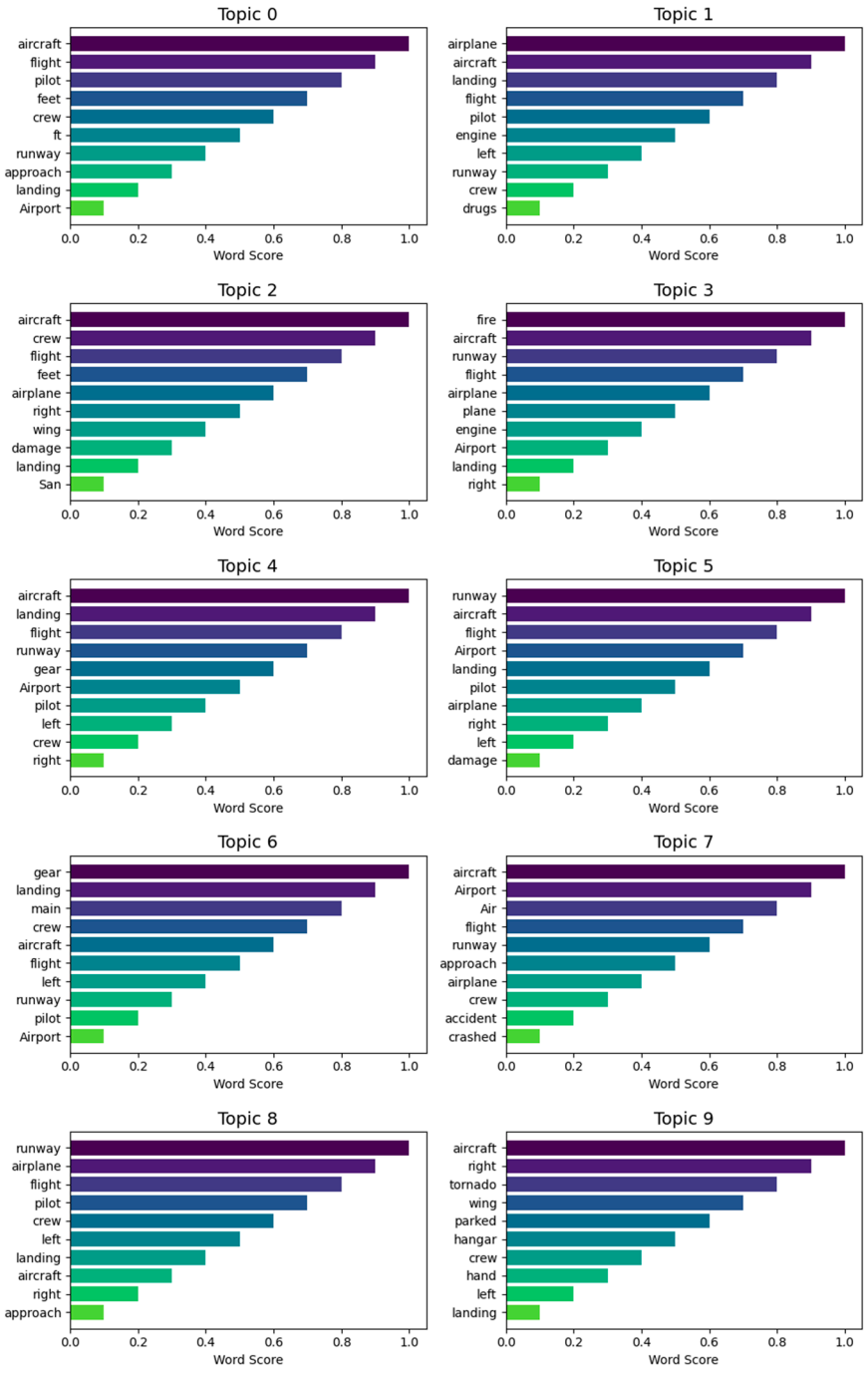

Figure 3. Top words for each topic chosen by the pLSA model.

Table 1. Top words from BERTopic and pLSA for each topic, and their associated theme.

Topic	BERTopic Top 10 Words	pLSA Top 10 Words	Theme / Single Word
0	caravan, grand, cessna, forced, simikot, near, airstrip, impacted, terrain, pilot	aircraft, flight, pilot, feet, crew, ft, runway, approach, landing, Airport	Small Aircraft and Flight Basics
1	illegal, venezuelan, drugs, venezuela, mexican, mexico, colombian, jet, guatemala, xb	airplane, aircraft, landing, flight, pilot, engine, left, runway, crew, drugs	Drug Trafficking and Flight Landing
2	otter, twin, servo, elevator, nancova, tourmente, dq, ononge, hinge, col	aircraft, crew, flight, feet, airplane, right, wing, damage, landing, San	Aircraft Parts and Flight Damage

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

3	fire, smoke, extinguished, parked, fireball, cargo, bottles, heat, emanating, rescue	fire, aircraft, runway, flight, airplane, plane, engine, Airport, landing, right	Fire Incident and Flight
4	caught, fire, canadair, erupted, repair, huatulco, hockey, arson, forced, providence	aircraft, landing, flight, runway, gear, Airport, pilot, left, crew, right	Fire Event and Landing Gear
5	tornado, blown, substantially, tune, storm, hangered, nashville, damaged, tennessee, struck	runway, aircraft, flight, Airport, landing, pilot, airplane, right, left, damage	Tornado Damage and Runway Incident
6	learjet, paso, toluca, mateo, olbia, iwakuni, mexico, cancn, vor, michelena	gear, landing, main, crew, aircraft, flight, left, runway, pilot, Airport	Jet and Airports and Gear and Emergency
7	bird, birds, flock, strike, windshield, geese, remains, roskilde, spar, multiple	aircraft, Airport, Air, flight, runway, approach, airplane, crew, accident, crashed	Bird Strike and Accidents
8	havana, cuba, bogot, medelln, rionegro, permission, carreo, tulcn, haiti, lamia	runway, airplane, flight, pilot, crew, left, landing, aircraft, right, approach	Latin America and Runway and Flight
9	medan, tower, supervisor, acted, rendani, indonesia, pk, controller, ende, jalaluddin	aircraft, right, tornado, wing, parked, hangar, crew, hand, left, landing	Air Traffic Control and Tornado Damage

Figure 4. Top words for each topic chosen by the BERTopic model.

4.2.2. Visualization and Interpretability

Figure 2 displays the Intertopic Distance Map generated by BERTopic using UMAP. Each circle represents a topic, where size indicates frequency and color denotes cluster affiliation. The D1 and D2 axes reflect semantic similarity between topics based on their vectorized embeddings. The map also demonstrates the effectiveness of BERTopic's topic reduction mechanism, which allows fine-tuning of topic granularity post hoc. In contrast, Figures 5 and 6 presents the top words for each topic chosen by BERTopic and pLSA respectively, highlighting the relative importance of each word within its assigned topic. Although pLSA offers statistically tight clusters, BERTopic's visualization facilitates better exploration and thematic understanding, especially for non-technical users.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

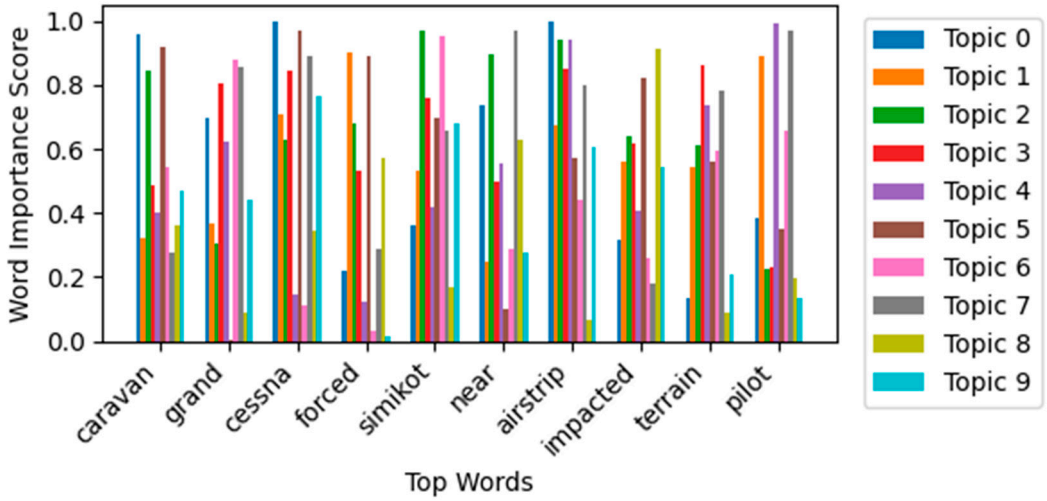

Figure 5. Top words for each topic chosen by BERTopic.

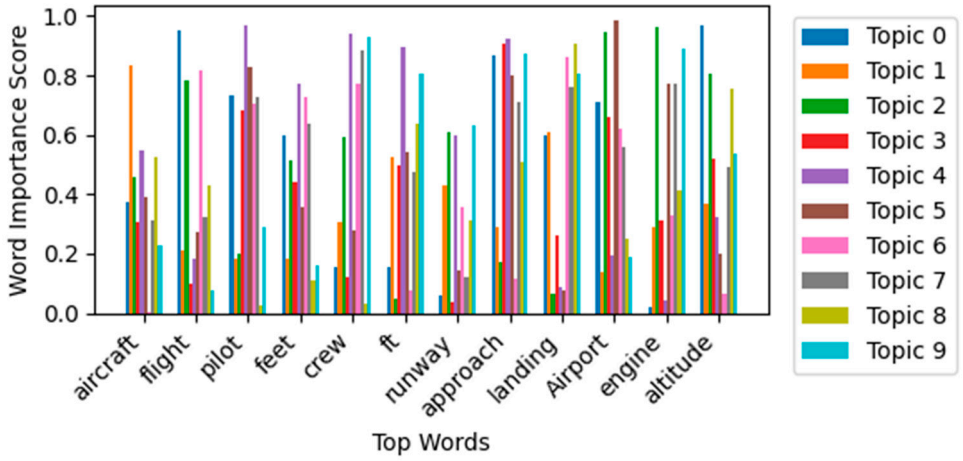

Figure 6. Top words for each topic chosen by PLSA.

4.3. Word Clouds and Topic Distribution

Beyond semantic clustering, this study employed word cloud visualizations to further explore the lexical distribution and salience of topic terms as shown in Figures 7 and 8. This offers a visual summary of the most prominent terms across topics. BERTopic's cloud displayed more differentiated and semantically tight terms, further supporting the interpretability advantage. Figure 9 shows the topic distribution from the pLSA model, where topic dominance and overlap can be inferred, it shows that the most prominent words were chosen by topic 2. BERTopic's distribution, though not shown here, displayed a more balanced and non-redundant topic separation.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

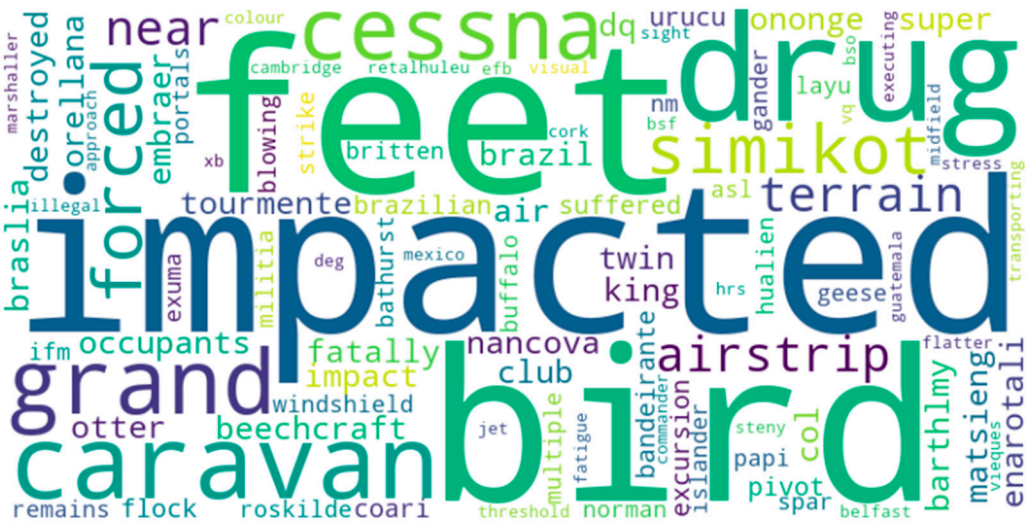

Figure 7. Wordcloud for BERTopic.

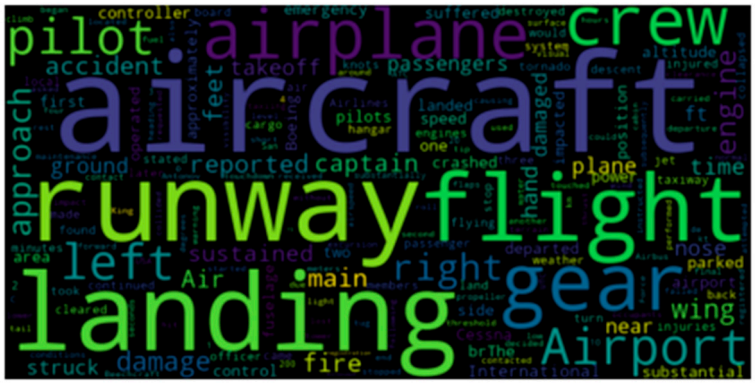

Figure 8. Wordcloud for pLSA.

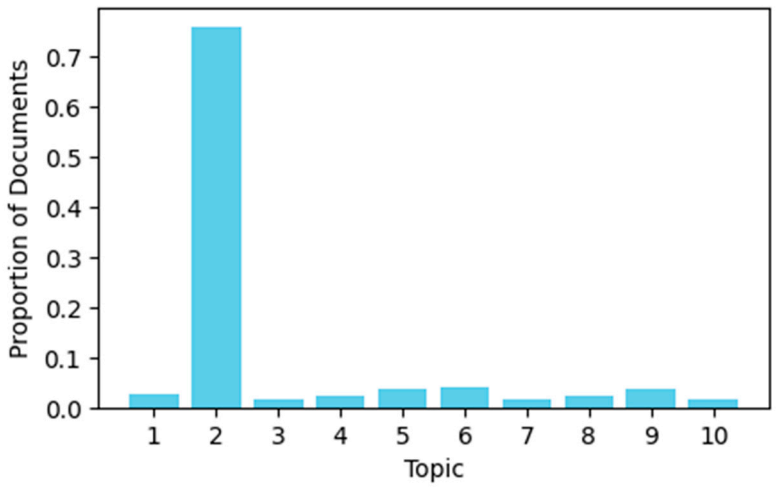

Figure 9. Topic word score for pLSA.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

4.4. Evaluation of Model Properties

Table 2 summarizes the observed strengths and weaknesses of the two models. While BERTopic demonstrated flexibility in topic resolution and clarity in thematic visualization, pLSA excelled in computational efficiency and statistical coherence.

Table 2. Comparison of model strengths and weaknesses.

Property	BERTopic	pLSA
Interpretability	High	Moderate
Coherence	0.531	0.7634
Perplexity	−4.532	−4.6237
Granularity Control	Adjustable	Fixed
Computational Cost	Higher	Lower
Visualization	Strong (via UMAP)	Limited

5. Ablation Study

To further understand the individual contributions of each modeling component, an ablation study was conducted on both pLSA and BERTopic models. The study systematically varied preprocessing steps, topic count, and dimensionality reduction techniques to evaluate their effect on performance metrics and interpretability. Interpretability outcomes from each variant were qualitatively assessed by aviation experts, reinforcing the practical relevance of model adjustments.

For the pLSA model, removing stop word filtering and lemmatization led to a 7.4% drop in coherence, affirming the importance of these preprocessing techniques. Additionally, varying the number of topics from 5 to 20 revealed an optimal range between 9 and 12 topics, beyond which topics became redundant or overly fragmented. Similarly, the use of TF-IDF versus raw term frequency showed negligible impact on perplexity but reduced coherence by 0.06.

In contrast, BERTopic's performance was more sensitive to changes in the embedding model. When Sentence-BERT embeddings were replaced with TF-IDF vectors, coherence dropped from 0.531 to 0.411. The choice of dimensionality reduction method was also crucial. Using PCA instead of UMAP reduced visual clarity and interpretability in the inter-topic distance map, reinforcing the effectiveness of UMAP for semantic clustering [21]. Adjusting the min_topic_size hyperparameter revealed that larger values improved coherence slightly but led to loss of granularity and omitted minority topics, which are essential in aviation safety data.

The higher alignment of BERTopic's outputs with known aviation patterns, as recognized by expert reviewers, underscores its practical applicability in real-world aviation safety investigations. These findings highlight the nuanced trade-offs between algorithm complexity, interpretability, and statistical performance, reinforcing the need to balance quantitative metrics with domain relevance.

5.1. Discussion

The incorporation of aviation experts during manual evaluation provided essential feedback on the practical relevance and clarity of the topics generated. Their insights were particularly valuable in validating themes such as Bird Strike Investigations and Helicopter Operations, which may not have been obvious from a purely statistical perspective. The results indicate distinct strengths and trade-offs between the two topic modeling approaches. While pLSA achieved higher coherence and lower perplexity, its topics often lacked semantic cohesion upon manual inspection. This suggests that statistical optimization alone does not guarantee meaningful interpretability, a critical requirement in aviation safety applications. BERTopic, although scoring lower in coherence, produced thematically dense and interpretable topics such as Bird Strike Investigations and Engine Failures, aligning more closely with aviation-specific incident patterns.

This difference can be attributed to BERTopic's integration of transformer-based embeddings and class-based TF-IDF (c-TF-IDF) for keyword extraction [28]. These mechanisms allow the model to capture deeper contextual relationships among terms, which pLSA's bag-of-words assumption

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

cannot fully account for. Furthermore, BERTopic's interactive inter-topic distance map enhances the model's usability, especially for stakeholders who benefit from visually intuitive representations of risk clusters.

Nevertheless, pLSA's higher coherence and simpler architecture suggest it remains viable for resource-constrained settings or when scalability is paramount. The results advocate for hybrid approaches or model stacking to leverage the interpretability of BERTopic with the statistical robustness of pLSA.

5.2. Limitations

Several limitations must be acknowledged in this study. First, the dataset is limited to publicly available reports from the ASN database, potentially excluding proprietary or more recent incident data. Second, topic labels were manually interpreted, which may introduce subjective bias. However, this was mitigated through cross-validation by multiple reviewers, including aviation experts, to ensure consistency and domain relevance. Additionally, while coherence and perplexity are standard metrics, they may not fully capture the domain-specific relevance of topics in aviation safety [32].

The BERTopic model's reliance on transformer embedding introduces a computational burden, making it less practical for real-time or low-resource environments. Furthermore, the pLSA model assumes independence among topics and does not incorporate word embeddings, which limits its semantic richness. Lastly, the models were not evaluated in downstream applications such as risk classification or incident prediction, which would further validate their practical utility.

6. Conclusions

This study presents a comparative analysis of two topic modeling approaches—pLSA and BERTopic on aviation safety incident narratives. The inclusion of domain experts in the evaluation process not only enhanced thematic validation but also demonstrated the value of expert-informed machine learning in safety-critical domains. While pLSA demonstrated superior statistical coherence and lower perplexity, BERTopic outperformed in thematic clarity and user-friendly visualization. The findings underscore the importance of balancing algorithmic performance with domain-specific interpretability, especially in critical sectors such as aviation.

Future work will focus on integrating domain-specific ontologies and external aviation taxonomies to enrich topic labeling. Additionally, combining topic modeling with supervised classification tasks could enable the identification of high-risk events in real time. Exploring more recent models like Top2Vec or embedding-powered LDA variants, and evaluating cross-lingual performance for global datasets, also represents a promising direction [33].

Lastly, we plan to extend this work by applying topic modeling to incident causality analysis, enabling safety analysts to proactively identify latent factors contributing to aviation accidents, thereby supporting data-driven decision-making and regulatory oversight.

Author Contributions: A.N.: conceptualization, methodology, software, data curation, validation, writing—original draft preparation, formal analysis, K.J. validation, writing—review and editing and U.T.: writing—review and editing, and G.W.: data collection, supervision, final draft.

Funding: This research received funding from the Tuition Fee Scholarship at UNSW.

Data Availability Statement: The data that support the findings of this study are publicly available from the Aviation Safety Network (ASN) at https://aviation-safety.net/. The dataset includes unstructured narrative reports and associated damage level classifications.

Conflicts of Interest: The authors declare no conflicts of interest.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Full Form
ASN	Aviation Safety Network
ATSB	Australian Transport Safety Bureau
BERTopic	Bidirectional Encoder Representations from Transformers Topic Modeling
DL	Deep Learning
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
ML	Machine Learning
NLP	Natural Language Processing
NTSB	National Transportation Safety Board
pLSA	Probabilistic Latent Semantic Analysis
TF-IDF	Term Frequency-Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection

References

1. A. Nanyonga and G. Wild, "Impact of Dataset Size & Data Source on Aviation Safety Incident Prediction Models with Natural Language Processing," in *2023 Global Conference on Information Technologies and Communications (GCITC)*, 2023, pp. 1-7: IEEE.
2. A. Nanyonga, K. Joiner, U. Turhan, and G. Wild, "Applications of natural language processing in aviation safety: A review and qualitative analysis," in *AIAA SCITECH 2025 Forum*, 2025, p. 2153.
3. A. Gupta and H. J. N. Fatima, "Topic modeling in healthcare: A survey study," vol. 20, no. 11, pp. 6214-6221, 2022.
4. A. J. Rawat, S. Ghildiyal, and A. K. J. I. J. E. T. T. Dixit, "Topic Modeling Techniques for Document Clustering and Analysis of Judicial Judgements," vol. 70, no. 11, pp. 163-169, 2022.
5. M. Apishev, S. Koltcov, O. Koltsova, S. Nikolenko, and K. Vorontsov, "Additive regularization for topic modeling in sociological studies of user-generated texts," in *Advances in Computational Intelligence: 15th Mexican International Conference on Artificial Intelligence, MICAI 2016, Cancún, Mexico, October 23–28, 2016, Proceedings, Part I 15*, 2017, pp. 169-184: Springer.
6. H. Axelborn and J. Berggren, "Topic Modeling for Customer Insights: A Comparative Analysis of LDA and BERTopic in Categorizing Customer Calls," ed, 2023.
7. A. Nanyonga, K. Joiner, U. Turhan, and G. Wild, "Does the Choice of Topic Modeling Technique Impact the Interpretation of Aviation Incident Reports? A Methodological Assessment," 2025.
8. T. Hofmann, "Probabilistic latent semantic analysis," in *UAI*, 1999, vol. 99, pp. 289-296.
9. Y. Mu, C. Dong, K. Bontcheva, and X. J. a. p. a. Song, "Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling," 2024.
10. J. A. dos Santos, T. I. Syed, M. C. Naldi, R. J. Campello, and J. J. I. T. o. B. D. Sander, "Hierarchical density-based clustering using MapReduce," vol. 7, no. 1, pp. 102-114, 2019.
11. M. Masseroli, D. Chicco, and P. Pinoli, "Probabilistic latent semantic analysis for prediction of gene ontology annotations," in *The 2012 international joint conference on neural networks (IJCNN)*, 2012, pp. 1-8: IEEE.
12. M. Wahabzada and K. Kersting, "Larger residuals, less work: Active document scheduling for latent Dirichlet allocation," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, 2011, pp. 475-490: Springer.
13. A. Nanyonga and G. J. a. p. a. Wild, "Analyzing Aviation Safety Narratives with LDA, NMF and PLSA: A Case Study Using Socrata Datasets," 2025.
14. J. Rusakovica, J. Hallinan, A. Wipat, and P. J. J. o. i. b. Zuliani, "Probabilistic latent semantic analysis applied to whole bacterial genomes identifies common genomic features," vol. 11, no. 2, pp. 93-105, 2014.
15. M. La Rosa, A. Fiannaca, R. Rizzo, and A. J. B. b. Urso, "Probabilistic topic modeling for the analysis and classification of genomic sequences," vol. 16, pp. 1-9, 2015.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2025 doi:10.20944/preprints202505.1509.v1

16. S. T. Dumais, "LSA and information retrieval: Getting back to basics," in *Handbook of latent semantic analysis*: Psychology Press, 2007, pp. 305-334.
17. N. C. J. T. D. S. Albanese, "Topic Modeling with LSA, pLSA, LDA, NMF, BERTopic, Top2Vec: a Comparison," vol. 19, no. 09, 2022.
18. S. Xu, Y. Wang, X. Cheng, and Q. Yang, "Thematic Identification Analysis of Equipment Quality Problems Based on the BERTopic Model," in *2024 6th Management Science Informatization and Economic Innovation Development Conference (MSIEID 2024)*, 2025, pp. 484-491: Atlantis Press.
19. H. Sibitenda *et al.*, "Extracting Semantic Topics about Development in Africa from Social Media," 2024.
20. A. Nanyonga, H. Wasswa, U. Turhan, K. Joiner, and G. Wild, "Comparative Analysis of Topic Modeling Techniques on ATSB Text Narratives Using Natural Language Processing," in *2024 3rd International Conference for Innovation in Technology (INOCON)*, 2024, pp. 1-7: IEEE.
21. L. McInnes, J. Healy, and J. J. a. p. a. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018.
22. Y. Kim and H. J. I. J. o. C. Kim, "An Analysis of Research Trends on the Metaverse Using BERTopic Modeling," vol. 19, no. 4, pp. 61-72, 2023.
23. W. Chen, F. Rabhi, W. Liao, and I. J. E. Al-Qudah, "Leveraging state-of-the-art topic modeling for news impact analysis on financial markets: a comparative study," vol. 12, no. 12, p. 2605, 2023.
24. R. Egger and J. J. F. i. s. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," vol. 7, p. 886498, 2022.
25. A. Nanyonga, H. Wasswa, U. Turhan, K. Joiner, and G. Wild, "Exploring Aviation Incident Narratives Using Topic Modeling and Clustering Techniques," in *2024 IEEE Region 10 Symposium (TENSYMP)*, 2024, pp. 1-6: IEEE.
26. A. Agovic, H. Shan, and A. Banerjee, "Analyzing Aviation Safety Reports: From Topic Modeling to Scalable Multi-Label Classification," in *CIDU*, 2010, pp. 83-97: Citeseer.
27. D. Gefen, J. E. Endicott, J. E. Fresneda, J. Miller, and K. R. J. C. o. t. A. f. I. S. Larsen, "A guide to text analysis with latent semantic analysis in R with annotated code: Studying online reviews and the stack exchange community," vol. 41, no. 1, p. 21, 2017.
28. M. J. a. p. a. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022.
29. R. IBRAIMOH, K. O. DEBRAH, and E. NWAMBUONWO, "Developing & Comparing Various Topic Modeling Algorithms on a Stack Overflow Dataset," 2024.
30. S. Deb and A. K. J. M. L. w. A. Chanda, "Comparative analysis of contextual and context-free embeddings in disaster prediction from Twitter data," vol. 7, p. 100253, 2022.
31. A. Hoyle, P. Goel, A. Hian-Cheong, D. Peskov, J. Boyd-Graber, and P. J. A. i. n. i. p. s. Resnik, "Is automated topic model evaluation broken? the incoherence of coherence," vol. 34, pp. 2018-2033, 2021.
32. M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399-408.
33. D. J. a. p. a. Angelov, "Top2vec: Distributed representations of topics," 2020.

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.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.