aerospace

MDPI

Article

AI-Driven Modeling of Near-Mid-Air Collisions Using Machine Learning and Natural Language Processing Techniques

Dothang Truong

School of Graduate Studies, Embry-Riddle Aeronautical University, Daytona Beach, FL 32114, USA; truongd@erau.edu

Abstract

As global airspace operations grow increasingly complex, the risk of near-mid-air collisions (NMACs) poses a persistent and critical challenge to aviation safety. Traditional collision-avoidance systems, while effective in many scenarios, are limited by rule-based logic and reliance on transponder data, particularly in environments featuring diverse aircraft types, unmanned aerial systems (UAS), and evolving urban air mobility platforms. This paper introduces a novel, integrative machine learning framework designed to analyze NMAC incidents using the rich, contextual information contained within the NASA Aviation Safety Reporting System (ASRS) database. The methodology is structured around three pillars: (1) natural language processing (NLP) techniques are applied to extract latent topics and semantic features from pilot and crew incident narratives; (2) cluster analysis is conducted on both textual and structured incident features to empirically define distinct typologies of NMAC events; and (3) supervised machine learning models are developed to predict pilot decision outcomes (evasive action vs. no action) based on integrated data sources. The analysis reveals seven operationally coherent topics that reflect communication demands, pattern geometry, visibility challenges, airspace transitions, and advisory-driven interactions. A four-cluster solution further distinguishes incident contexts ranging from tower-directed approaches to general aviation pattern and cruise operations. The Random Forest model produces the strongest predictive performance, with topic-based indicators, miss distance, altitude, and operating rule emerging as influential features. The results show that narrative semantics provide measurable signals of coordination load and acquisition difficulty, and that integrating text with structured variables enhances the prediction of maneuvering decisions in NMAC situations. These findings highlight opportunities to strengthen radio practice, manage pattern spacing, improve mixed equipage awareness, and refine alerting in short-range airport area encounters.

Keywords: aviation safety; near mid air collision (NMAC); machine learning; natural language processing (NLP); airspace operations; pilot evasive action

Check for updates

Academic Editor: Graham Wild

Received: 30 November 2025
Revised: 5 January 2026
Accepted: 9 January 2026
Published: 12 January 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The safety of the global airspace system remains one of the most critical concerns in modern aviation, particularly as operational environments become increasingly complex as traffic volumes grow, with the integration of unmanned aircraft systems (UASs) or drones and heterogeneous aircraft operations. Among the most severe indicators of this system complexity are near-mid-air collisions (NMACs), events in which two aircraft come into dangerously close proximity, typically defined in U.S. regulations as less than 500 ft or when a pilot or flight crew member reports that a collision hazard existed [1]. NMACs

represent the final layer between routine operations and catastrophic collision, where seconds determine whether an evasive maneuver prevents tragedy or a fatal event ensues.

Recent accidents and incidents underscore the urgent relevance of this issue. On 29 January 2025, the mid-air collision at the Ronald Reagan Washington National Airport (DCA) between a PSA Airlines Bombardier CRJ-700 and a U.S. Army UH-60 Black Hawk helicopter claimed 67 lives, highlighting the persistent vulnerability of shared airspace in mixed helicopter–fixed-wing corridors [2]. Subsequent investigations revealed that more than 15,000 near-miss encounters between helicopters and airplanes had been reported in the same region within three years before the crash—an "intolerable risk," according to the NTSB [3]. Around the same period, multiple near-mid-air incidents occurred in the United States, such as the August 2025 Goodyear, Arizona, event involving a Van's RV-7 and a Diamond DA-40 near the traffic-pattern altitude of Goodyear Airport [4]. These incidents demonstrate that NMAC risks are neither isolated nor confined to commercial aviation; rather, they are pervasive across general aviation, rotorcraft, business jets, and emerging UAS operations [5].

Although traditional separation assurance and collision-avoidance systems, such as the Traffic Alert and Collision Avoidance System (TCAS) and the newer Airborne Collision Avoidance System (ACAS X), have significantly reduced mid-air collision risks, their rule-based logic and reliance on transponder data may be insufficient under the dynamic and mixed-traffic conditions of today's airspace [6,7]. The proliferation of UAS, urban air mobility platforms, and increasingly complex flight operations introduces new variables, ranging from communication latency to pilot situational awareness, that challenge existing systems and call for data-driven analytic frameworks capable of capturing emerging risk patterns.

Unstructured narrative data, such as those collected in the NASA Aviation Safety Reporting System (ASRS), contain a wealth of contextual information describing pilot perceptions, environmental factors, operational constraints, and decision processes [8]. Yet, much of this valuable data remains underutilized by traditional statistical approaches. Recent advances in machine learning (ML) and natural language processing (NLP) offer a unique opportunity to extract latent patterns from these narratives, enabling deeper insight into the precursors and dynamics of NMAC incidents [9]. By analyzing text alongside structured metadata, it becomes possible to identify thematic clusters of incidents and even predict critical decision outcomes, such as whether a pilot takes evasive action when faced with an imminent conflict.

This study, therefore, proposes a novel, integrative machine learning framework for NMAC analysis, structured around three methodological pillars. First, NLP techniques are applied to ASRS narratives to extract latent topics and semantic features describing operational, environmental, and human factors contexts. Second, cluster analysis is employed on both textual and structured features to identify homogeneous incident groups that represent distinct NMAC typologies. Third, supervised ML models, including decision trees, random forests, gradient boosting, and neural networks, are trained to predict pilot decision outcomes (evasive action vs. no action). The combined framework facilitates both pattern discovery and predictive insight, bridging data analytics and human-factor understanding.

More specifically, this research seeks to answer three research questions:

1. What latent topics can be extracted from ASRS NMAC narratives using NLP, and how do they expand understanding of contributing factors?
2. How can cluster analysis of these features and structured data generate empirically grounded profiles of NMAC incidents that differ in operational and human factors dimensions?

3. Which machine learning algorithms most effectively predict the pilot decision outcome, and what features, narrative, contextual, or environmental, contribute most to that prediction?

The contributions of this study are threefold. First, it demonstrates that integrating NLP-based topic modeling with structured data can uncover hidden precursors of NMACs beyond standard taxonomic classifications. Second, it implements predictive modeling of pilot decision-making, a rarely explored dimension in NMAC research, by linking thematic features and contextual variables to the likelihood of evasive action. Third, it prioritizes model interpretability and domain relevance through transparent feature-importance metrics, cluster profiling, and narrative-topic mapping, thereby enabling translation of ML insights into practical safety management.

In other words, this study moves beyond algorithmic novelty to deliver actionable safety intelligence. By combining narrative-driven analytics, clustering, and prediction within one unified framework, it responds to the dual challenges of increasing airspace congestion and the integration of diverse vehicle types. Ultimately, the goal is not merely to predict whether a pilot acts, but to reveal *why* pilots make those decisions under varying contexts, an understanding essential to training, human–machine interface design, and proactive air-traffic management.

The remainder of this paper is organized as follows: The next section provides a critical review of prior studies in machine-learning-based aviation safety, NLP analysis of safety narratives, and predictive modeling of pilot behavior, highlighting research gaps this work seeks to address. The subsequent sections present the methodology, results, discussion, and implications for both research and practice.

2. Literature Review and Contributions of This Study

2.1. Previous Work

2.1.1. Studies on Near-Mid-Air Collisions (NMACs)

Near-mid-air collisions (NMACs) have long been recognized as a critical safety issue in both commercial and general aviation, as they represent the final barrier before a catastrophic event. Historically, research on NMACs has concentrated on air traffic control (ATC) procedures, pilot situational awareness, and collision avoidance systems such as the Traffic Alert and Collision Avoidance System (TCAS) and Airborne Collision Avoidance System (ACAS X).

Early work by Kochenderfer et al. [6] at MIT Lincoln Laboratory developed probabilistic encounter models to estimate mid-air collision risks, laying the foundation for formal, quantitative modeling of encounter dynamics. These models relied primarily on radar and surveillance data, using Monte Carlo simulations to quantify risk and evaluate TCAS performance. Using a Colored Petri Net simulation framework, Tang et al. [10] showed that TCAS II can produce induced collisions under certain multi-aircraft and high-density encounter conditions, revealing that advisory timing and interaction effects can limit the system's effectiveness in unsegregated airspace.

Subsequent studies have highlighted several important contributing factors to NMACs. Stroeve [11] demonstrated that surveillance sensor uncertainty, variability in pilot response, and encounter geometry shape NMAC risk by limiting the ability of TCAS II and ACAS Xa to issue timely and effective advisories. Figuet et al. [12] showed that collision risk increases in dense or structurally complex airspace where rapid closure rates, unexpected trajectory deviations, and insufficient realized separation develop more frequently. Haselein et al. [13] extended this perspective by demonstrating that NMAC events are influenced not only by encounter geometry but also by human performance factors such as crew configuration, fatigue indicators, and operational context extracted from ASRS narratives. In complemen-

tary work, Nordlund and Gustafsson [14] found that uncertainty in aircraft state estimation, including sensor noise, trajectory prediction error, and measurement latency, amplifies conflict detection errors and increases the likelihood of delayed or ineffective avoidance maneuvers. Tuncal [15] suggests that existing research often fails to account for unique flight dynamics specific to UAVs as opposed to traditional aircraft, a differentiation crucial for accurately predicting collisions in non-segregated airspace. Finally, Vera et al. [16] demonstrated that a monocular camera plus an optical-flow–based obstacle-detection algorithm can successfully detect incoming airborne traffic under simulated encounter scenarios, highlighting that visual motion cues, independent of transponder or radar data, can be leveraged for mid-air collision avoidance.

Taken together, these studies indicate that NMACs arise from the interaction of human, environmental, and operational conditions rather than from a single causal factor. However, the literature still lacks a comprehensive investigation of how these influences combine in nonlinear ways across diverse operational contexts, which motivates further work in integrated modeling approaches.

2.1.2. Machine Learning Approaches in NMAC Research

Recent years have witnessed a paradigm shift toward data-driven safety analytics, with machine learning (ML) being increasingly employed to detect latent safety risks and predict incident likelihood [17,18]. Berges et al. [19] indicate that adversarial machine learning techniques, specifically gradient-based perturbation methods and neural-network–driven spoofing models, can manipulate TCAS surveillance inputs to induce unsafe resolution advisories, thereby increasing the likelihood of NMAC scenarios. Zhou et al. [20] developed a deep learning framework using convolutional and recurrent neural network architectures to identify and predict hazardous aircraft states from large-scale flight-data recorder datasets. Their results demonstrated that deep learning models can capture complex, nonlinear relationships among flight variables that traditional threshold-based or statistical monitoring methods fail to detect.

More recent research has expanded the use of ML to real-world operational and reporting systems. Haselein et al. [13] applied multiple ML models to the NASA ASRS database to predict and explain NMAC risk patterns. Their results underscored the potential of data-driven models to enhance risk awareness and improve real-time hazard detection. Similarly, Vinogradov et al. [21] developed a hybrid ML framework integrating UAV remote identification systems with air traffic data, demonstrating improved prediction of loss-of-separation events in mixed manned–unmanned airspace.

Parallel developments have occurred in natural language processing (NLP) applied to aviation safety narratives. Yang and Huang [22] conducted a systematic review of NLP applications in aviation, highlighting the ability of topic modeling and text mining to extract latent safety themes from narrative reports such as those in the ASRS. Paradis et al. [23] advanced this by developing *Kaona*, a deep-learning platform for mining ASRS narratives to support hazard curation and pattern detection. Despite these advances, most narrative-based approaches remain limited to topic discovery or categorization, with few linking unstructured narrative insights to predictive modeling or behavioral outcomes such as pilot evasive action.

Overall, previous research demonstrates two critical trends. First, machine learning has proven highly effective in capturing nonlinear, high-dimensional risk factors associated with NMACs and loss-of-separation events. Second, NLP provides a powerful means to mine human factors and contextual insights from narrative data. Yet these two analytical paradigms, structured ML and unstructured NLP, have seldom been integrated into a

unified framework that simultaneously clusters incident typologies and predicts pilot decisions in near-collision contexts.

2.2. Research Gaps

Despite significant progress in machine learning and safety analytics, several key research gaps persist that constrain comprehensive understanding and predictive capability in NMAC studies.

Limited Integration of Structured and Unstructured Data. Most prior studies on NMACs and loss-of-separation events have relied on structured data sources, such as radar tracks, flight parameters, or simulation outputs, while underutilizing the rich qualitative narratives contained in the ASRS. These narratives capture pilot cognition, communication dynamics, and situational factors that structured data cannot convey. The absence of integrative modeling that combines quantitative and qualitative features limits the depth of insight into incident causality and decision processes.

Underexplored Use of NLP for NMAC Pattern Discovery. Although recent research [22,23] has applied NLP to aviation safety reports, few studies have focused specifically on NMAC-related narratives. Consequently, the potential for topic modeling and semantic clustering to reveal latent NMAC groups, based on shared linguistic or contextual patterns, remains largely unexplored. This gap inhibits the ability to identify emergent operational risk clusters that transcend conventional categorical labels.

Lack of Predictive Models for Pilot Decision-Making. The decision to perform evasive action is a critical determinant of NMAC outcomes, yet the literature provides little empirical modeling of this decision behavior. Existing studies [13,21] have predicted NMAC occurrence or proximity, but not *pilot response*. Odisho et al. [24] focused on pilots' decision prediction, but mainly in unstable approach situations rather than NMACs. Understanding the factors that drive a pilot's decision to maneuver or not, based on operational, environmental, or narrative data, is essential for improving safety training and alerting systems.

Insufficient Focus on Explainability and Operational Interpretability. While many ML-based aviation safety studies achieve strong predictive performance, few address model transparency or interpretability. Tuncal [15] emphasizes that air traffic controllers face challenges in managing UAV risks, partly due to the unfamiliar flight characteristics of UAVs and the limited interpretability of collision prediction models. The lack of explainable AI in safety-critical domains limits real-world adoption. Without interpretable outputs, such as feature importance or topic-to-decision mapping, insights remain inaccessible to practitioners responsible for operational safety [25].

Emerging Complexity in Hybrid Airspace Operations. The integration of UAS, advanced air mobility (AAM), and mixed-traffic corridors introduces new dynamics that traditional NMAC models, developed for manned operations, fail to capture. Studies such as Vinogradov et al. [21] show promise but remain early-stage. A need exists for predictive models that account for hybrid operational contexts and evolving traffic management paradigms.

These research gaps directly align with the purpose of the present study, which seeks to combine NLP-driven topic extraction with machine-learning-based clustering and prediction to better understand and forecast pilot decision-making in NMAC incidents using ASRS data.

2.3. Contributions of This Study

In response to the identified gaps, the present study contributes to the advancement of aviation safety research in three principal ways:

Integration of NLP and Machine Learning for NMAC Analysis. This study introduces an analytical framework that integrates NLP-derived thematic topics with structured ASRS metadata to analyze NMAC incidents. Through topic modeling and semantic feature extraction, this study captures latent themes, operational contexts, and behavioral cues that traditional numerical approaches overlook.

Clustering and Typology Discovery of NMAC Incidents. Using unsupervised machine learning, this study develops empirically grounded clusters of NMAC events, revealing distinct incident groups characterized by shared thematic, operational, and environmental patterns. This clustering approach transcends standard ASRS taxonomy, providing a data-driven lens for categorizing near-collision risk.

Predictive Modeling of Pilot Decision Outcomes. By implementing supervised machine learning algorithms, this study models the pilot's decision to perform evasive action as a function of both structured and narrative-derived features. This predictive framework quantifies the influence of contextual and linguistic cues on pilot behavior, offering insights relevant to training, situational awareness, and decision-support system design. This research incorporates interpretable ML techniques to identify and visualize the most influential predictors of pilot evasive action. These interpretable results enhance trust and applicability in operational settings such as air traffic management and flight training, supporting data-driven decision-making consistent with Safety Management System (SMS) principles.

By integrating topic detection, clustering, and prediction, this study advances proactive safety analytics, moving from retrospective investigation to predictive and prescriptive risk management. The findings contribute to the broader goal of developing intelligent, explainable, and adaptive safety systems capable of supporting both manned and unmanned operations in complex airspace.

Overall, existing research has established the feasibility of applying machine learning to aviation safety and demonstrated the value of narrative analysis for understanding human factors. However, a unified framework that synthesizes these approaches to discover, classify, and predict behavioral responses in NMAC incidents remains lacking. The present study addresses this gap by applying NLP, clustering, and predictive modeling to NASA ASRS data, thereby contributing new theoretical, methodological, and practical insights for enhancing safety and decision-making in modern airspace systems.

3. Methodology

3.1. Research Design

This study employs the mixed-methods approach, consisting of both quantitative and qualitative safety data. It uses an archival, data-driven research design integrating natural language processing (NLP), unsupervised clustering, and supervised machine learning (ML) techniques to analyze NMAC incidents reported in the U.S. national airspace system. This research follows a three-phase analytical sequence: (1) *topic modeling* using Latent Dirichlet Allocation (LDA)of safety narratives, (2) *two-step cluster analysis* to derive data-driven incident typologies, and (3) *predictive modeling* to forecast pilot decision outcomes. This sequential design enables both pattern discovery and behavioral prediction, bridging descriptive narrative mining and inferential modeling.

3.2. Data Source and Collection

3.2.1. Data Source

The dataset was obtained from the NASA Aviation Safety Reporting System (ASRS), a publicly accessible repository of de-identified, voluntary safety reports submitted by pilots, controllers, and maintenance personnel. The ASRS database provides structured

event metadata (e.g., aircraft type, phase of flight, weather conditions) and unstructured narrative text describing the incident context, operational dynamics, and human decision processes [26].

3.2.2. Data Extraction and Filtering

To capture all relevant NMAC incidents, the ASRS database was queried using the following parameters:

- **Event Type:** Conflict—NMAC;
- **Timeline:** 1988–2025;
- **Search Scope:** All U.S. airspace classes and aircraft types.

After filtering and cleaning for duplicate and incomplete entries, the final dataset contained 13,111 validated reports. Each report included a detailed set of structured metadata describing aircraft configuration and operational context, as well as free-text fields documenting pilot actions. Table 1 presents the variables selected for this study. The variable descriptions are derived from the ASRS Coding Taxonomy [27].

Table 1. Description of variables used in the analyses.

Variables	Descriptions
Date	Year and month of the incident (YYYYMM)
Local Time Of Day	Time interval: 0001–0600, 0601–1200, 1201–1800, 1801–2400
State Reference	U.S. state where the incident occurred
Relative Position.Angle.Radial	Bearing/radial from a reference point
Relative Position.Distance.Nautical Miles	Distance from reference point in nautical miles
Altitude.AGL.Single Value	Altitude above ground level
Altitude.MSL.Single Value	Altitude above mean sea level
Flight Conditions	Weather flight category (VMC, IMC, Mixed, Marginal, etc.)
Flight Plan	Type of flight plan: IFR, VFR, DVFR, SVFR, or none
Weather Elements/Visibility	Specific weather elements: clouds, fog, visibility, hail, haze/smoke, icing, rain, etc.
Light	Lighting condition (Dawn, Daylight, Dusk, Night)
Ceiling	Ceiling condition: Ceiling height or CLR.
RVR.Single Value	Runway visual range
Aircraft 1 Operator	Operator type of Aircraft 1
Aircraft 1 Operating Under 14 CFR Part	Regulatory part for Aircraft 1
Aircraft 1 Flight Plan	Flight plan for Aircraft 1
Aircraft 1 Mission	Primary flight phase of Aircraft 1
A1_Phase_Primary_8Phase	Mission purpose of Aircraft 1
Aircraft 1 Airspace	Airspace for Aircraft 1
Aircraft 2 Operator	Operator type of Aircraft 2
Aircraft 2 Operating Under 14 CFR Part	Regulatory part for Aircraft 2
Aircraft 2 Flight Plan	Flight plan for Aircraft 2
A2_Phase_Primary_8Phase	Primary flight phase of Aircraft 2
Aircraft 2 Mission	Mission purpose of Aircraft 2
Aircraft 2 Airspace	Airspace for Aircraft 2

Table 1. *Cont.*

Variables	Descriptions
Miss Distance Horizontal	Estimated horizontal miss distance between aircraft (if a conflict occurred)
Miss Distance Vertical	Estimated vertical miss distance between aircraft
Contributing Factors	Key factors affecting the event: human factors, weather, airport, airspace, chart/publication issues, equipment/tooling, procedures, staffing, automation, etc.
Primary Problem	Main causal category: aircraft, airport, airspace structure, company policy, equipment/tooling, weather, chart/publication, software/automation, procedure, manuals, logbook, ambiguous, etc.
Results	Documented event outcome
Narratives	Reporter's free-text description of the event

3.2.3. Target Variable: Pilot Decision Outcome

The binary target variable represented whether the pilot executed **evasive action (1)** or **no evasive action (0)** in response to a near-collision threat. Classification was determined based on the "Results" variable in the dataset and on manual verification of a representative sample to ensure coding reliability.

3.3. Topic Modeling

3.3.1. Text Preparation and Normalization

Each narrative was normalized through a standard NLP preprocessing pipeline that included lowercasing, tokenization, punctuation and stop-word removal, and lemmatization. Aviation-specific terms and acronyms (e.g., *ATC*, *TCAS*, *IFR*, *VFR*) were preserved via a custom domain dictionary. Tokens occurring in fewer than five documents or in more than 95% of reports were excluded to reduce noise and improve model sparsity.

3.3.2. Topic Modeling via Latent Dirichlet Allocation (LDA)

Rationale for LDA choice. To extract latent semantic themes from the narrative corpus, the Latent Dirichlet Allocation (LDA) algorithm was employed. LDA was selected because it produces probabilistic, document-level topic weights that map directly to continuous predictors suitable for clustering and supervised modeling, and because its topics are readily interpretable by domain experts—an important design criterion for safety-critical, practitioner-facing applications. Although modern neural embedding approaches (e.g., contextual transformers) offer greater representational richness, LDA's transparency and its ability to support topic visualization and labeling were prioritized in this study [28].

Conceptually, LDA models each document (narrative) as a probabilistic mixture of latent topics, where each topic is defined as a probability distribution over words [29–32]. The process involved:

1. Vectorization: For the Latent Dirichlet Allocation (LDA) topic model, a document-term count matrix (word counts per document) was created after standard preprocessing (lowercasing, tokenization, punctuation removal, lemmatization, domain dictionary preservation of aviation acronyms) and frequency filtering (tokens occurring in <5 documents or >95% of documents were removed). A TF–IDF representation was computed separately for exploratory similarity checks and for some classifier variants, but the LDA estimation used the count matrix to satisfy LDA's multinomial assumptions. We evaluated candidate k values using coherence, perplexity diag-

nostics, and manual interpretability checks and selected k = 7 as the best balance of coherence and operational interpretability [33].

2. Model Estimation: Multiple LDA models were tested using topic numbers ranging from three to ten; the optimal model was selected based on coherence score, perplexity, and manual interpretability.
3. Topic Selection: The final LDA model identified dominant topics, each representing recurrent thematic dimensions in NMAC incidents along with corresponding keywords. Each topic will receive an appropriate label reflecting the theme of that topic.
4. Feature Creation: Topic probability distributions (θ-values) for each narrative were extracted and used as continuous predictor variables in the clustering and prediction phases.

This topic modeling step operationalized qualitative narratives into structured, high-dimensional semantic vectors, enabling subsequent quantitative analysis.

3.3.3. Integration with Structured Metadata

Narrative-derived topic features were merged with structured ASRS variables, such as aircraft type, flight phase, meteorological conditions, airspace class, and time of day, to create a unified dataset for clustering and modeling. All numeric features were standardized using z-score normalization to eliminate scale bias.

3.4. Analytical Procedures

3.4.1. Two-Step Cluster Analysis

To identify latent NMAC typologies, a two-step clustering approach combining hierarchical and partitioning methods was employed. In the first step, hierarchical clustering using Ward's minimum variance method was used to assess the natural structure of the data and determine the preliminary number of clusters based on the dendrogram and agglomeration coefficients. In the second step, K-means clustering was applied, using the optimal cluster number from the hierarchical stage as the initial seed to refine partition boundaries and minimize within-cluster variance [25,34].

The resulting clusters represented distinct NMAC typologies characterized by both thematic and operational attributes. Cluster quality was evaluated using the silhouette coefficient and between-cluster separation indices.

3.4.2. Predictive Modeling of Pilot Decision Outcomes

To predict whether pilots executed evasive maneuvers, a suite of supervised machine learning algorithms was developed and compared.

- Binary Logistic Regression;
- Artificial Neural Network (Multilayer Perceptron);
- Decision Tree;
- Gradient Boosting;
- Random Forest.

Because the target variable was imbalanced—evasive maneuvers occurring more frequently than non-evasive outcomes—the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set to generate synthetic examples of the minority class. This step mitigated bias toward the majority class, improved sensitivity, and ensured that all algorithms were trained on a more balanced representation of pilot decision outcomes [25].

The dataset was partitioned into three samples: training (60%), testing (20%), and validation (20%). The training sample was used to train the models. The testing sample was used to test the model, and the validation sample was used to finally validate the model. The partition and validation process is to ensure the model reliability and validity, minimize

overfitting, and ensure the generalizability of the final model [25]. All algorithms were implemented in SPSS Modeler Premium 26. Hyperparameters were optimized using grid search and five-fold cross-validation to ensure generalizable performance. We acknowledge that random splits do not directly address temporal drift over 1988–2025; this choice was made to evaluate baseline generalization across a heterogeneous corpus. Temporal cross-validation (time-based folds) is recommended for future work to evaluate forecasting to later years or to assess temporal stability of topic prevalence and model coefficients.

3.4.3. Model Performance Evaluation

Model reliability was assessed by comparing performance metrics between the training, testing, and validation samples to identify potential overfitting or underfitting. Stable and consistent metrics across the three samples were interpreted as evidence of model reliability and robustness. Model validity was evaluated using a comprehensive set of statistical and classification metrics [25]:

- Accuracy—proportion of correctly classified observations;
- Misclassification rate—complement of accuracy;
- Recall (sensitivity)— true positive rate (probability of correctly detecting evasive actions);
- Specificity—true negative rate (probability of correctly detecting non-evasive outcomes);
- F1-Score—harmonic mean of precision and recall;
- Gini Index—measure of discriminative power ($2 \times \text{AUC} - 1$);
- Lift—measure of model improvement over random classification;
- Receiver Operating Characteristic (ROC) Curve—area under the curve (AUC) used as the primary comparative indicator of overall performance.

Models were compared across all these metrics to determine the best-performing algorithm. The results were also examined through feature importance analysis to identify which narrative topics and operational conditions most strongly influenced the likelihood of evasive action.

Overall, the combination of NLP, clustering, and predictive modeling provides a robust, interpretable, and replicable methodological structure for understanding the patterns and determinants of pilot behavior in NMAC events.

4. Results

4.1. Descriptive Statistics

The dataset comprised 13,111 NMAC reports submitted to the NASA Aviation Safety Reporting System (ASRS) between August 1988 and May 2025, after data cleaning. These records represent voluntary safety submissions from pilots and other flight personnel across U.S. states and territories. Each report included structured information (e.g., operation type, phase of flight, flight rules, altitude, and environmental conditions) as well as a narrative describing the event circumstances. The combined structure provided a comprehensive overview of operational contexts and situational factors associated with NMAC incidents (Table 2).

Reporter and Operation Characteristics

Most reports originated from pilots, consistent with ASRS submission patterns. Regarding operation type, Part 91 general aviation dominated the dataset with 6587 reports (50.2%), followed by Part 121 air carrier operations (2534, 19.3%) and Part 135 charter operations (670, 5.1%). Smaller categories included Part 119, Part 137, and Part 107 (unmanned aircraft systems). The distribution for the second aircraft in each event showed a similar pattern: Part 91 (4462 reports), Part 121 (736), and Part 135 (250). This pattern indicates that NMAC incidents most frequently involve general aviation aircraft, either among themselves or in interactions with commercial carriers.

Flight and Environmental Conditions

A majority of NMACs occurred during terminal and transition phases, where traffic density and pilot workload are greatest. The initial approach phase accounted for 3326 incidents (26.2%), followed by cruise (3099, 23.6%), descent (1306, 10%), climb (1266, 9.7%), and final approach (1166, 8.9%). Additional occurrences were recorded during initial climb (1003, 7.6%), landing (713, 5.4%), and takeoff or launch (572, 4.4%).

With respect to flight rules, for Aircraft 1, instrument flight rules (IFR) were reported in 1861 cases (14.19%), while visual flight rules (VFR) appeared in 2557 (19.5%). About 4263 reports (32.5%) did not specify a flight plan, which likely reflects missing data in narrative-based submissions. These results suggest that while NMAC risk exists across both controlled and uncontrolled environments, general aviation and VFR operations remain more frequently represented.

Weather and lighting conditions followed similar patterns to prior ASRS findings. Visual meteorological conditions (VMC) prevailed in 10,812 incidents (82.5%), whereas instrument meteorological conditions (IMC) occurred in 384 (2.9%) and mixed or marginal weather in fewer than 7%. Daylight operations accounted for the majority (10,326 reports, 78.8%), with night (839, 6.4%), dusk (532, 4.1%), and dawn (63, 0.5%) comprising smaller proportions. These findings indicate that NMACs are most likely to occur in clear weather and daytime conditions—periods of high visibility but also high traffic activity.

Evasive Action Outcomes

Among all records, 8812 reports (67.2%) indicated that evasive action was taken, while 4299 (32.8%) reported no evasive action. This binary outcome reflects the operational decision-making context of each event and serves as the target variable in subsequent predictive modeling. The high proportion of positive evasive responses suggests that in most cases, crews recognized the conflict early enough to maneuver, although a notable proportion of late or absent reactions underscores ongoing challenges in situational awareness and communication.

Temporal and Operational Trends

Longitudinal analysis shows a distinct pattern over time. Following relatively higher counts in the late 1980s and early 1990s, the introduction and widespread adoption of the Traffic Collision Avoidance System (TCAS) in the mid-1990s corresponded with a gradual decline in NMAC reports through the early 2000s. Beginning around 2015, however, the rapid growth of unmanned aircraft systems (UAS), particularly small recreational drones operating in the National Airspace System, contributed to a renewed increase in NMAC reports. In addition, afternoon operations (1200–1800 local time) showed the highest reporting frequency, aligning with peak approach and training traffic periods. Collectively, these temporal and operational patterns reinforce that most midair conflicts occur during routine, high-density flight windows rather than under unusual or adverse conditions.

Table 2. Descriptive characteristics of NMAC reports.

Variable	Frequency	Percent (%)
Operation Type (Aircraft 1)		
Part 91	6587	50.2
Part 121	2534	19.3
Part 135	670	5.1
Other/Unknown	139	1.1
Operation Type (Aircraft 2)		
Part 91	4462	34.0
Part 121	736	5.6

Table 2. *Cont.*

Variable	Frequency	Percent (%)
Part 135	250	1.9
Other/Unknown	496	3.8
Flight Phase (Aircraft 1)		
Initial Approach	3326	26.2
Cruise	3099	23.6
Descent	1306	10.0
Climb	1266	9.7
Final Approach	1166	8.9
Initial Climb	1003	7.6
Landing	713	5.4
Takeoff/Launch	572	4.4
Flight Phase (Aircraft 2)		
Cruise	3336	25.44
Initial Approach	2224	16.96
Climb	942	7.18
Final Approach	723	5.51
Landing	611	4.66
Initial Climb	596	4.55
Descent	586	4.47
Takeoff/Launch	459	3.50
Flight Plan (Aircraft 1)		
IFR	1861	14.19
VFR	2557	19.5
Not Specified	4263	32.5
Flight Plan (Aircraft 2)		
IFR	1276	14.19
VFR	2019	15.40
Not Specified	7591	57.9
Flight Conditions		
VMC	10,812	82.5
IMC	384	2.9
Mixed/Marginal	651	5.0
Light Conditions		
Daylight	10,326	78.8
Night	839	6.4
Dusk/Dawn	595	4.5
Evasive Action		
Yes	8812	67.2
No	4299	32.8

Note. The table reports the most frequent categories; percentages do not sum to 100%. IFR = Instrument Flight Rules; VFR = Visual Flight Rules; VMC = Visual Meteorological Conditions; IMC = Instrument Meteorological Conditions.

4.2. Topic Modeling Results

A total of 13,111 NMAC narrative reports extracted from the NASA ASRS database (1988–2025) were processed using the Latent Dirichlet Allocation (LDA) model to identify latent thematic patterns. The k-selection diagnostics support a seven-topic solution (Figure 1). As shown in the plot, coherence remains relatively stable across values of k but reaches one of its better points at k = 7, while topic diversity stays high and comparable across all tested values. Perplexity, plotted on the secondary axis, decreases from k = 5 to k = 7 and then levels off, showing little meaningful improvement beyond k = 7. The convergence of these metrics indicates that k = 7 provides the most balanced structure: it maintains strong topic quality and diversity while avoiding the unnecessary fragmentation that occurs when additional topics are introduced. This selection is also consistent with foundational principles of LDA, which seek a value of k that maximizes model fit without over-specifying the latent structure, ensuring that topics remain probabilistically coherent, stable, and interpretable.

The inter-topic distance map corroborates this choice: seven centroids occupy distinct regions with limited overlap, and the small residual intersections appear along conceptually adjacent traffic-pattern themes rather than across unrelated areas, which is expected for narrative aviation data (Figure 2). Finally, the topic prevalence chart shows a healthy distribution: one operational theme is prominent but not monopolizing, and the remaining topics capture meaningful variance rather than noise (Figure 3). Together, these diagnostics justify selecting k = 7.

The seven topics identified by the LDA model reflect complementary facets of operational communication and traffic management, illustrating how pilots and controllers coordinate to maintain situational awareness and mitigate conflict risk in shared airspace (Table 3).

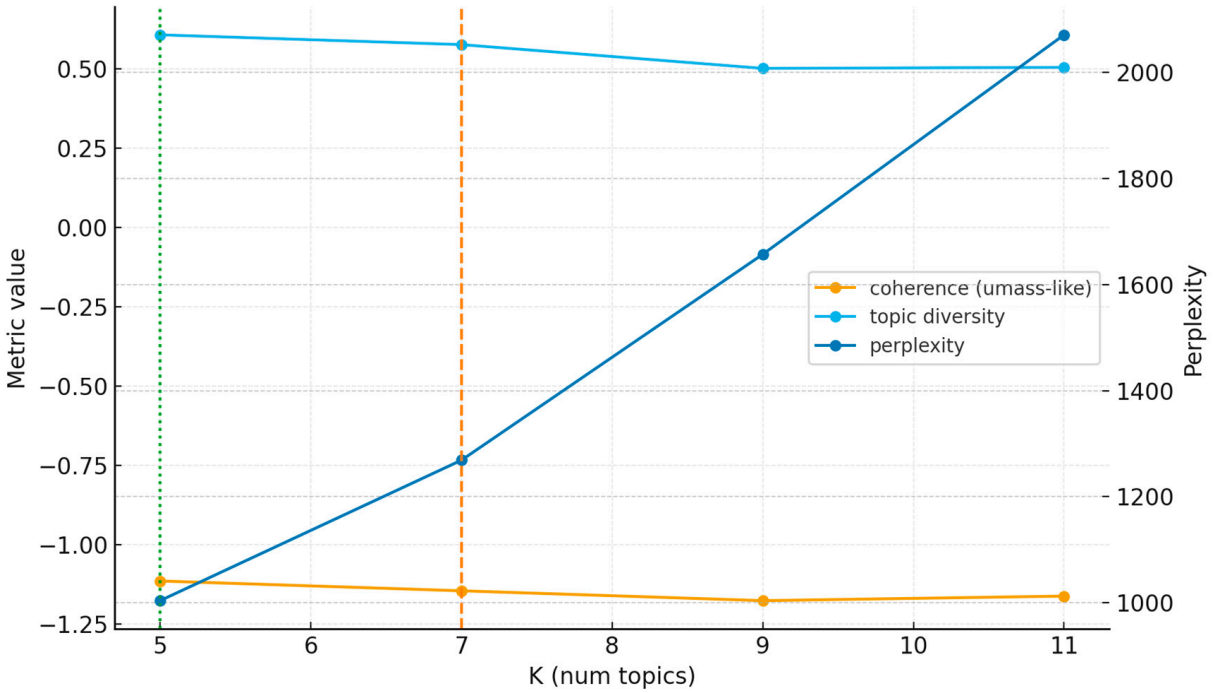


Figure 1. K selection diagnostics.

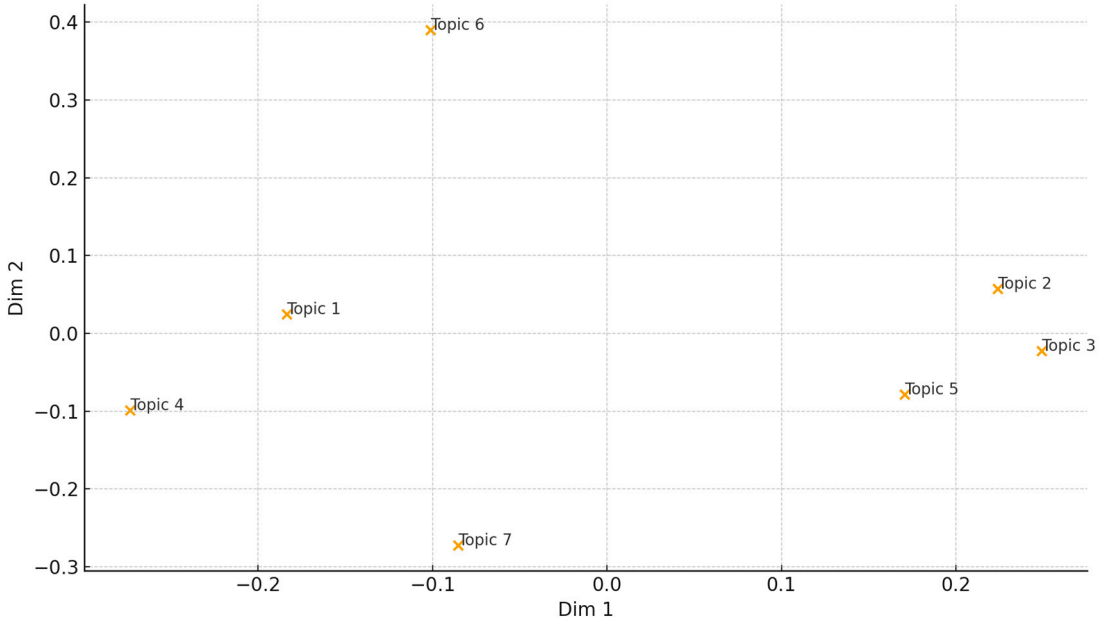


Figure 2. Inter-topic distance map.

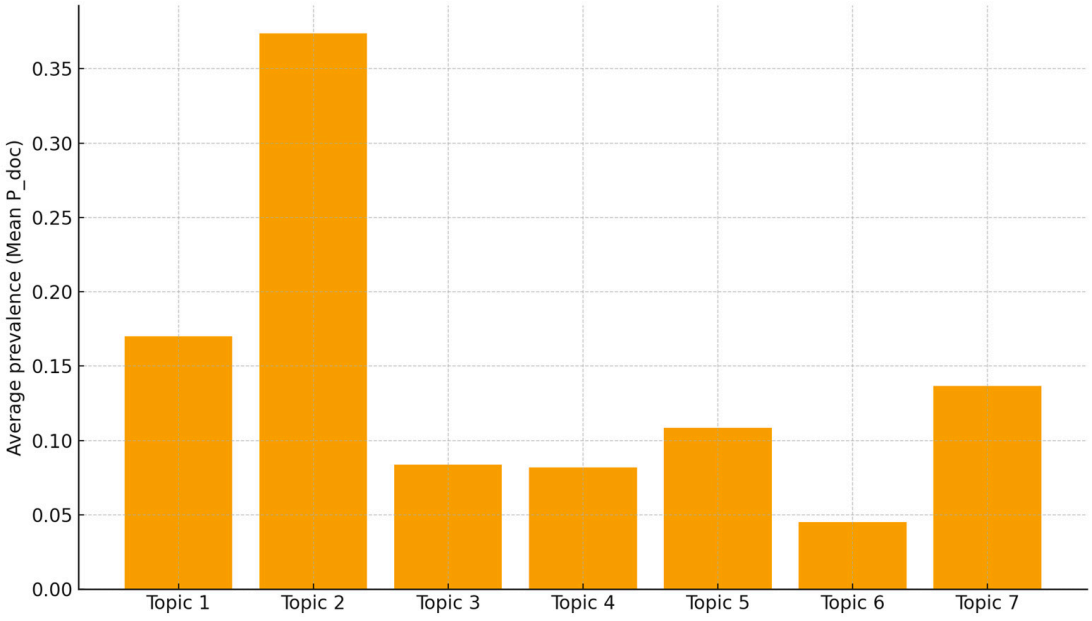


Figure 3. Topic prevalence across NMAC reports.

- ***Topic 1: Traffic-Pattern Operations and Training Flights*** captures radio calls and maneuvering within the VFR pattern, particularly involving student pilots, left/right traffic, CTAF announcements, and spatial positioning around the runway environment.
- ***Topic 2: Tower Communications, Clearances, and Approach Sequencing*** reflects structured interactions with tower and approach controllers, including traffic advisories, runway assignments, clearances, and sequencing during arrival and departure.
- ***Topic 3: Visual Sightings and See-and-Avoid Encounters*** centers on helicopter, drone, glider, and unidentified-aircraft sightings, where pilots report visually detecting other traffic, passing in close proximity, and occasionally conducting evasive actions.
- ***Topic 4: Tower–Runway Coordination and Helicopter Integration*** highlights runway operations directed by tower controllers, including takeoff and landing clearances,

helicopter traffic mixing with fixed-wing flows, and controller-issued instructions for maintaining separation.

- ***Topic 5: Airspace Structure, Class Transitions, and Radar Contact*** represents navigation through or near controlled airspace, including VFR/IFR class transitions, radar identification issues, and communication challenges associated with entering or exiting various airspace classes.
- ***Topic 6: Final Approach Geometry and Base-to-Final Conflicts*** emphasizes pattern geometry, especially short final, base-to-final turns, downwind entries, and spacing issues that emerge when multiple aircraft converge toward the runway environment.
- ***Topic 7: TCAS/ATC Traffic Advisories and Altitude/Climb–Descent Management*** reflects advisory-driven conflict detection, including traffic alerts, TCAS events, altitude adjustments, heading changes, and climb–descent profiles during approach and departure.

Table 3. Topic descriptions.

Topics	Label	Top 10 Items	Average Corpus Prevalence (Mean P_doc)
1	Traffic-Pattern Operations and Training Flights	pattern (0.022), downwind (0.017), student (0.012), runway (0.015), radio (0.011), left (0.010), ctaf (0.006), traffic (0.013), xx (0.006), right (0.009)	0.1700
2	Tower Communications, Clearances, and Approach Sequencing	tfc (0.022), twr (0.014), apch (0.014), rwy (0.015), ctlr (0.009), mi (0.008), tcasii (0.007), sma (0.006), clred (0.006), approx (0.007)	0.3700
3	Visual Sightings and See-and-Avoid Encounters	heli (0.016), drone (0.009), glider (0.007), white (0.006), lights (0.007), wing (0.010), passed (0.010), evasive (0.008), flying (0.010), appeared (0.007)	0.0836
4	Tower–Runway Interactions and Helicopter Operations	tower (0.038), runway (0.029), helicopter (0.012), cleared (0.010), controller (0.010), traffic (0.019), right (0.013), approach (0.012), landing (0.007), takeoff (0.005)	0.082
5	Airspace Structure, Class Transitions, and Radar Contact Issues	airspace (0.011), area (0.012), vfr (0.012), radar (0.009), class (0.007), ifr (0.006), class airspace (0.004), near (0.005), air (0.004), contact (0.005)	0.1086
6	Final Approach Geometry and Short-Final/Base-to-Final Conflicts	final (0.059), rwy (0.070), base (0.036), land (0.016), final rwy (0.007), short (0.011), short final (0.007), base final (0.007), downwind (0.016), landed (0.009)	0.0451
7	TCAS/ATC Traffic Advisories and Altitude/Climb–Descent Management	traffic (0.028), tcas (0.013), atc (0.015), feet (0.014), approach (0.014), climb (0.010), altitude (0.011), ra (0.011), heading (0.006), descent (0.005)	0.1366

Accordingly, these topics form a coherent structure of operational activities reflected in NMAC reports. Topics 2 and 4 capture tower-directed activity, including runway sequencing, clearances, and helicopter integration, while Topics 1, 5, and 6 represent VFR pattern dynamics, airspace structure, and approach-path geometry. Topics 3 and 7 focus on conflict detection - Topic 3 through visual see-and-avoid encounters involving helicopters,

drones, and gliders, and Topic 7 through TCAS alerts and ATC-issued traffic advisories that drive altitude and heading adjustments.

The inter-topic distance map shows a coherent structure in which Topics 2, 3, and 5 cluster together, reflecting their shared emphasis on tower communications, airspace navigation, and controller–pilot coordination. Topics 1 and 4 lie near each other in the central region, consistent with their common focus on VFR pattern operations and runway–tower interactions, while Topics 6 and 7 appear more separated due to their distinct emphasis on approach geometry and TCAS/ATC traffic advisories, respectively. Overall, the spatial layout supports the semantic relationships among topics, with related operational themes positioned closely and advisory-driven or altitude-management themes positioned further apart.

Token-salience profiles exhibit smooth β decay, indicating that each topic is semantically coherent rather than dominated by any single high-weight term. The prevalence distribution is also face-valid: tower and approach sequencing (Topic 2) appears most frequently, but the remaining themes collectively account for the majority of the corpus, providing broad coverage across traffic-pattern operations, airspace navigation, visual encounters, runway interactions, and advisory-driven avoidance events.

Overall, the seven-topic solution is both statistically justifiable and operationally meaningful. It captures the layered communication dynamics that underpin aviation safety—from structured air traffic control environments and airspace-navigation procedures to pattern coordination and real-time conflict avoidance, making it well suited for predictive analyses and the development of targeted safety interventions.

4.3. Cluster Analysis Results

4.3.1. Cluster Formation and Validation

A two-step cluster analysis was performed using topic-probability distributions of the seven topics found from the previous steps. The hierarchical clustering step applied Ward's linkage and Euclidean distance measures to determine grouping patterns [25,34]. The dendrogram and agglomeration schedule suggested a four-cluster solution, which was validated by the silhouette coefficient (0.63) and Calinski–Harabasz index (745.2) (Figure 4).

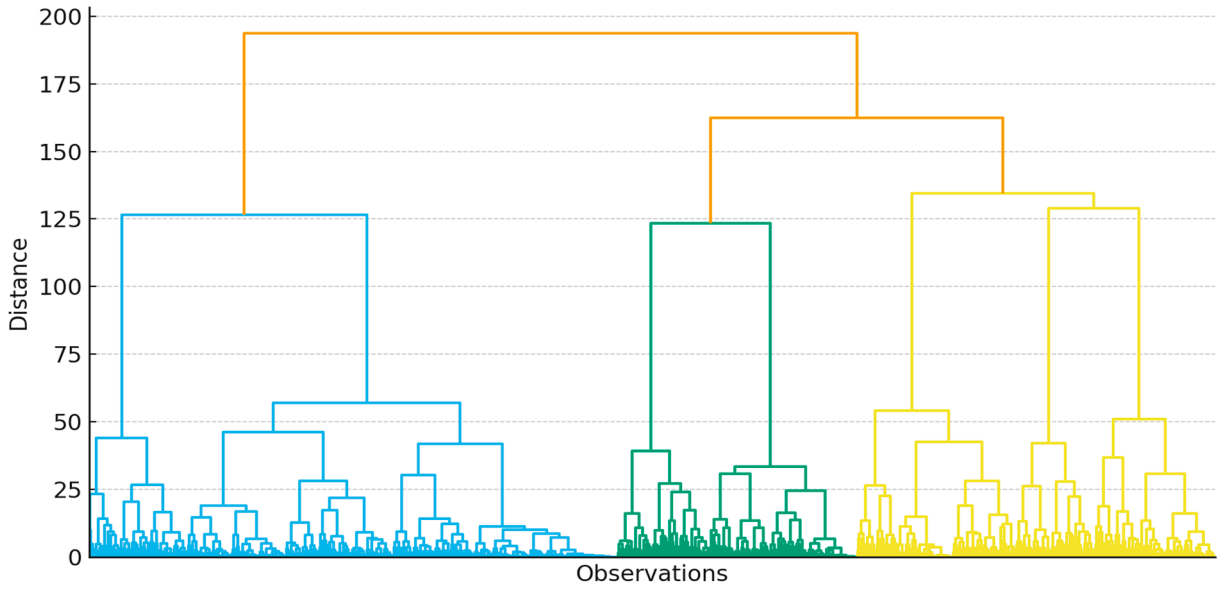


Figure 4. Dendrogram.

The second step refined the solution using K-means partitioning, initialized with centroids from the hierarchical stage. The resulting cluster memberships were stable across multiple random seeds, and inter-cluster separation exceeded 0.40 for all dimensions, confirming satisfactory discrimination.

The results of the cluster analysis revealed four distinct groups (Cluster 1, Cluster 2, Cluster 3, and Cluster 4) with a highly uneven distribution of observations (Figure 5). Cluster 1 constituted the majority of the dataset, representing 42.7% of all observations, making it the largest identified cluster. In contrast, Cluster 3 was determined to be the smallest group, accounting for only 13.8% of the data. The remaining two clusters, Cluster 2 and Cluster 4, showed relatively similar sizes, with Cluster 2 comprising 23.0% and Cluster 4 comprising 20.5%. Overall, the size disparity highlights the dominance of Cluster 1, which contains over three times the number of observations compared to the smallest group, Cluster 3.

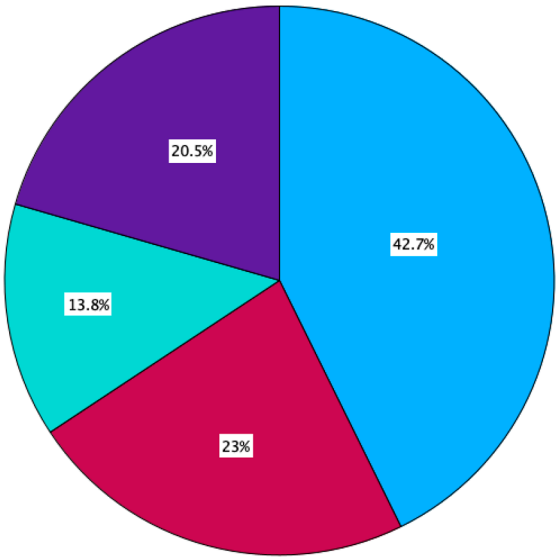


Figure 5. Cluster distribution.

The centroid-distance heatmap shows clear, non-trivial separation among all four clusters: no pair of centroids is "near-zero" (i.e., there is no obviously redundant pair to merge), and there is not a singleton far from the rest that would argue for an extra split (Figure 6). Distances are reasonably balanced across pairs, indicating four comparably distinct groups rather than two or three that dominate the structure. Together with the earlier validity evidence (dendrogram cut at four, silhouette = 0.63, Calinski–Harabasz = 745.2, and per-feature separation > 0.40), the inter-cluster distances support retaining a four-cluster solution as both interpretable and well separated.

The centroid analysis of topic probabilities showed clear thematic distinctions among the four clusters (Figure 7). Cluster C1 was dominated by Topic 2 (0.74), reflecting structured, tower-directed arrival operations. Cluster C2 emphasized Topic 1 (0.56) with some Topic 4 (0.25), indicating personal or training flights maneuvering in the pattern and interacting with tower instructions. Cluster C3 centered on Topic 7 (0.69), consistent with en-route or transitional phases where altitude changes and traffic alerts are more common. Finally, Cluster C4 showed a mixed profile with Topic 4 (0.25) and Topic 2 (0.08), representing general aviation operations that involve both runway–tower interactions and localized sequencing. Overall, these centroid patterns confirm that each cluster reflects a distinct operational context and communication profile within the NMAC narratives.

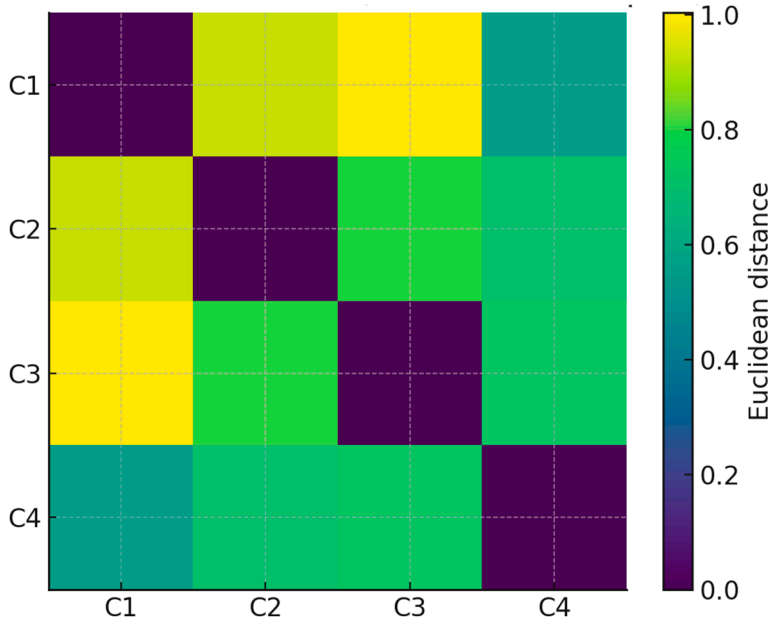


Figure 6. Inter-cluster distance map.

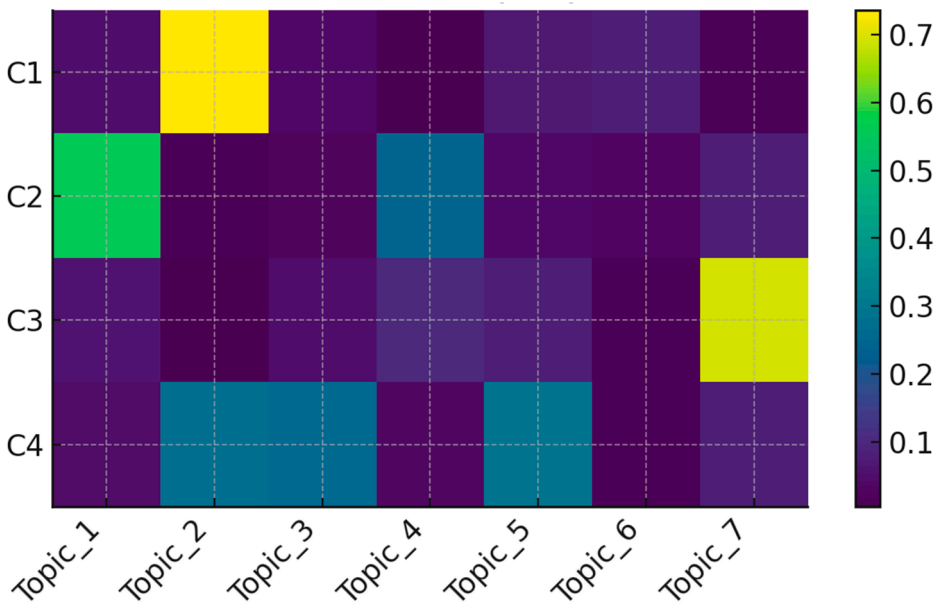


Figure 7. Cluster centroid on topic probabilities.

4.3.2. Cluster Profiling and Characteristics

Cluster profiling was conducted to create detailed, descriptive narratives for the groups (clusters) found during a cluster analysis, allowing us to understand the unique characteristics and conditions that define each NMAC incident typology [25,34]. These descriptive profiles help to translate the statistical groupings into actionable aviation safety insights, such as developing targeted risk mitigation strategies or tailored air traffic control procedures.

Table 4 shows the cluster profiling results. Variables that truly separate clusters include phase of flight and operational context (operator type, 14 CFR Part, flight plan), which show the cleanest between-cluster contrasts. Primary cause is mostly human factors across the board (52–87%); so, it is not a strong discriminator here. Time of day concentrates in the afternoon (1201–1800) for all clusters (~49–50%); so, it does not differentiate.

Table 4. Clustering profile.

Cluster ID	Cluster Name	Phase of Flight	Operation Type	Primary Cause	14 CFR Part	Flight Plan	Time of Day
1	Carrier Approach/IFR	Initial Approach (36%)	Air Carrier (45%)	Human Factors (87%)	Unknown (38%)	IFR (58%)	1201–1800 (49%)
2	Personal Part 91/Final Approach	Final Approach (26%)	Personal (49%)	Human Factors (72%)	Part 91 (89%)	None (51%)	1201–1800 (50%)
3	Carrier Cruise/IFR	Cruise (23%)	Air Carrier (41%)	Human Factors (52%)	Part 91 (50%)	IFR (63%)	1201–1800 (49%)
4	Personal Cruise/No Flight Plan	Cruise (51%)	Personal (36%)	Human Factors (75%)	Part 91 (46%)	None (38%)	1201–1800 (50%)

Based on these profiling results, we can interpret the characteristics of each cluster regarding NMAC incidents. This interpretation allows us to assign a descriptive name to each cluster that accurately reflects the dominant conditions and patterns of that group.

(1) Cluster C1—Carrier Approach/IFR
- Phase: Initial approach is the single largest category (36%).
- Operations: Air Carrier is dominant (45%); IFR plans are common (58%).
- 14 CFR Part: Many records have unknown reporting; no single 14 CFR Part dominates.
- Interpretation: Carrier operations concentrated around the approach group under instrument rules. This cluster likely reflects structured, procedure-driven operations near terminal areas.

(2) Cluster C2—Personal Part 91/Final Approach
- Phase: Final approach is the largest group (26%).
- Operations: Personal flying dominates (49%), overwhelmingly Part 91 (89%), with no flight plan recorded most often (51%).
- Interpretation: Predominantly personal GA operations in late descent/final approach, often without a flight plan. This cluster suggests pattern-entry/arrival contexts in GA.

(3) Cluster C3—Carrier Cruise/IFR (mixed Part 91/121 reporting)
- Phase: Cruise holds the lead (23%).
- Operations: Air Carrier has the largest share (41%), IFR is common (63%). 14 CFR Part skews Part 91 in the data (50%), implying some reporting/labeling heterogeneity or mixed operations.
- Interpretation: Carrier-heavy operations in en-route groups under IFR. The mixed 14 CFR Part suggests that some en-route encounters are captured under differing reporting conventions.

(4) Cluster C4—Personal Cruise/No Flight Plan
- Phase: Clearly Cruise-heavy (51%).
- Operations: Personal leads (36%), Part 91 is most frequent (46%); no flight plan is common (38%).
- Interpretation: General aviation cruise scenarios, often without a filed plan. This group typifies en-route GA contexts with less formalized procedural structure.

The profiling analyses show that the four topic-based clusters differ primarily by phase of flight and operational context. C1 (Carrier Approach/IFR) concentrates on initial-

approach groups in carrier operations with predominantly IFR flight plans. C2 (Personal Part 91/Final Approach) is driven by personal Part 91 flying during final approach and frequently without a flight plan. C3 (Carrier Cruise/IFR) reflects carrier-heavy, en-route operations under IFR (with some reporting heterogeneity across 14 CFR Parts), while C4 (Personal Cruise/No Flight Plan) captures general-aviation cruise operations, often without a flight plan. In contrast, human factors dominate the primary-cause field across clusters, and afternoon operations are common across the board; so, these variables provide limited discrimination. Overall, the profiling confirms that the clusters are meaningfully distinct in their operational setting and phase of flight, supporting the interpretability of the four-cluster solution.

Cluster size and implications. Cluster 1 comprises the largest share of reports (~42.7%). This dominance likely reflects two phenomena: (1) the operational fact that tower-sequenced approach operations generate frequent interactions and reporting, and (2) reporting heterogeneity in ASRS (e.g., higher reporting from commercial/training environments in terminal areas). The large size of Cluster 1, therefore, signals both a genuine concentration of NMAC reports in approach/tower contexts and a surveillance/reporting bias that should be accounted for when prioritizing interventions. Smaller clusters (e.g., Cluster 3) may represent lower-volume but operationally important contexts (such as en-route advisory events) and should not be neglected in safety planning. We recommend monitoring cluster prevalence over time to detect shifting risk profiles.

4.3.3. Cross-Cluster Variance

ANOVA tests on the mean topic weights across clusters yielded significant between-group differences ($F = 17{,}580.2$, $p < 0.001$ for Topic_2; $F = 17{,}544.0$, $p < 0.001$ for Topic_7), confirming that the clusters represent statistically distinct patterns rather than arbitrary groupings. Pairwise post hoc comparisons (Welch t-tests with Bonferroni adjustment) further indicated that Topic_2 is significantly higher in Cluster C1 than in C2, C3, and C4 (all adjusted $p < 0.05$), supporting the interpretive label we assigned to C1 (Carrier Approach/IFR).

4.4. Predictive Modeling Results

4.4.1. Model Training and Comparison

Five supervised ML algorithms were trained to predict the binary outcome of pilot evasive action (1) versus no evasive action (0): Binary Logistic Regression, Decision Tree (C5.0), Random Forest, Gradient Boosting Machine, and Bayesian Networks. Data were randomly split into 60% training, 20% test, and 20% validation samples. Due to the imbalanced dataset, the SMOTE approach was used to oversample the training data to ensure a balance between the two classes and minimize overfitting. Each model was optimized using five-fold cross-validation, and hyperparameters were tuned via grid search. Reliability was verified by comparing model metrics between training and validation samples; differences of less than 2% in accuracy and F1-score were considered acceptable.

The model comparison shows that Random Forest dominates the comparison across both accuracy and discrimination (Table 5). It attains the highest accuracy (0.879) and the lowest misclassification rate (0.121), while pairing strong recall (0.921) with solid specificity (0.837), which yields the best F1 score (0.884). Discrimination is also superior, with the largest Gini (0.924) and ROC AUC (0.962). Lift at 1.694 indicates that cases flagged positive by Random Forest are about 1.69 times as likely to be truly positive as a random case at the sample prevalence. Gradient Boosting is a distant second (accuracy 0.816, F1 0.8167, AUC 0.891, lift 1.629), while the single tree, Bayesian Networks, and Logistic Regression fall further behind. Given the safety objective to maximize detection without overwhelming

false alarms, Random Forest provides the best balance and the most flexible threshold tuning due to its superior ranking power and is selected as the champion model.

Table 5. Model comparison.

Model	Accuracy	Misclassification Rate	Specificity	Recall	F1	Gini Index	ROC AUC	Lift
Random Forest	0.879	0.121	0.837	0.921	0.884	0.924	0.962	1.694
Gradient Boosting	0.816	0.184	0.7971	0.836	0.816	0.782	0.891	1.629
Decision Tree (C5.0)	0.802	0.198	0.794	0.81	0.804	0.704	0.852	1.591
Bayesian Networks	0.723	0.277	0.686	0.76	0.733	0.604	0.802	1.577
Logistic regression	0.698	0.302	0.675	0.721	0.705	0.542	0.771	1.535

4.4.2. Champion Model Performance

Random Forest was selected as the champion model. The detailed configurations of this model are as follows:

Basic Settings:

- Number of Trees: 1000
- Maximum tree depth: 100
- Minimum leaf node size: 1
- Number of features to use for splitting: auto

Advanced Settings:

- Use bootstrap samples when building trees: Checked
- Use out-of-bag samples to estimate the generalization accuracy: Checked
- Use Extremely Randomized Trees: Checked
- Enable parallel model building: Checked
- Hyper-Parameter Optimization:
  - Target: 0.01
  - Max iterations: 1000
  - Max Evaluations: 300

Table 6 presents the model performance among the three samples. The model shows strong reliability, with nearly identical results across the three samples, indicating no overfitting issues. Accuracy is 0.883 on training, 0.879 on testing, and 0.883 on validation, a spread of about 0.4 percentage points. Discrimination is equally stable, with ROC AUC of 0.963, 0.962, and 0.962 and corresponding Gini values of 0.926, 0.924, and 0.925. Error profiles are consistent as well: specificity is 0.846, 0.837, and 0.852; recall is 0.921, 0.921, and 0.916; F1 is about 0.888, 0.884, and 0.887. Lift is also steady at roughly 1.71, 1.69, and 1.74. These small and symmetric differences indicate that the model generalizes well without signs of overfitting or instability.

The model's validity is also confirmed. The accuracy is about 88 percent on both the testing and validation sets, with misclassification around 11 to 12 percent, which is considered acceptable. This is supported by strong discrimination performance, with ROC AUC near 0.962 and Gini near 0.925, and a balanced confusion profile, where recall is about 0.92 and specificity about 0.84 to 0.85, yielding F1 near 0.88. Together, these results

demonstrate that the model is both reliable across samples and valid for predicting the target in this dataset.

Table 6. Random Forest model performance among three samples.

Samples	Accuracy	Misclassification	Specificity	Recall	F1	Gini Index	ROC AUC	Lift
Training	0.883	0.117	0.846	0.921	0.888	0.926	0.963	1.708
Testing	0.879	0.121	0.837	0.921	0.884	0.924	0.962	1.694
Validation	0.883	0.117	0.852	0.916	0.887	0.925	0.962	1.742

4.4.3. Feature Importance Analysis

Figure 8 shows the feature importance based on the results of the Random Forest model. The chart helps us better understand the key factors contributing to predicting the pilot decision during NMAC incidents.

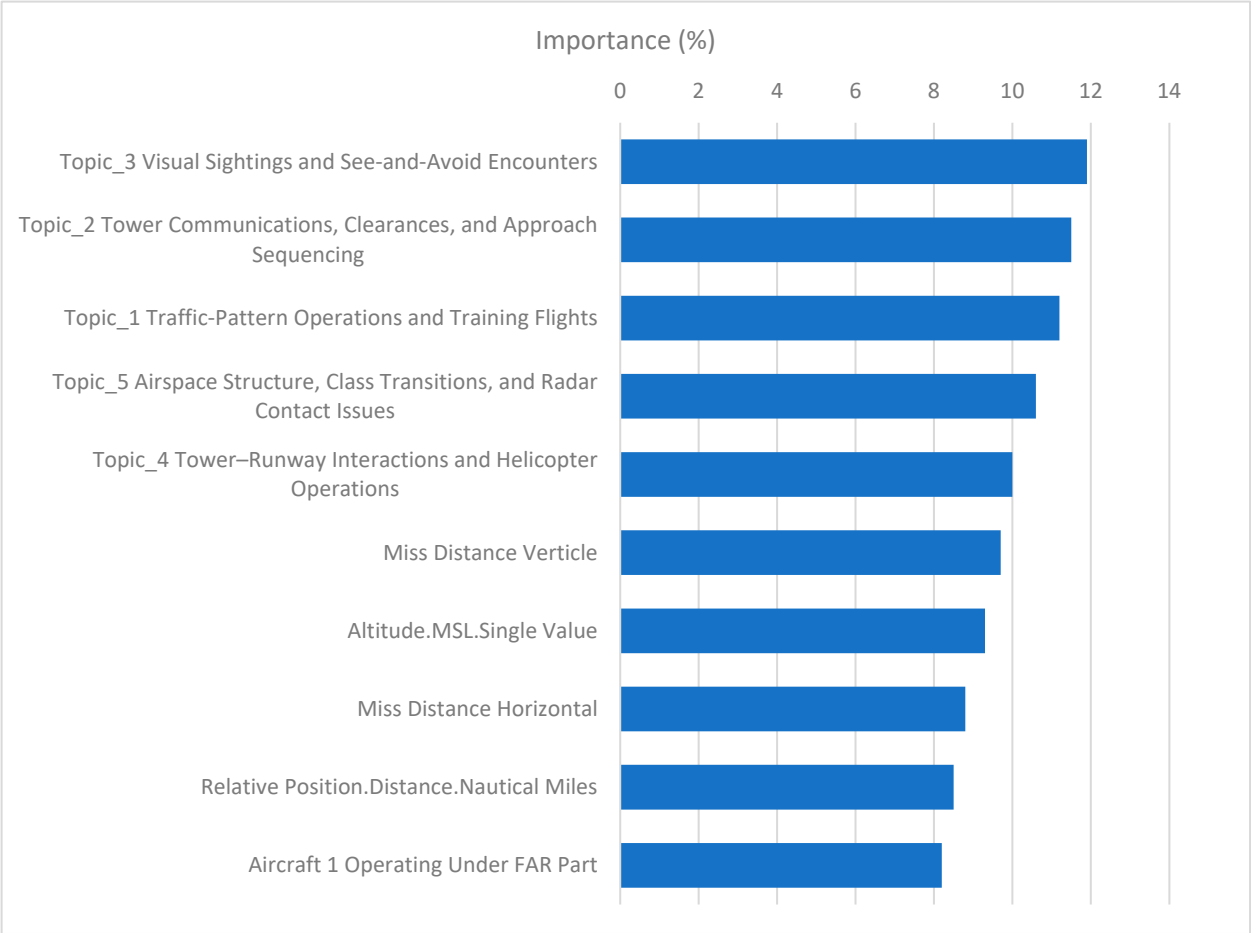


Figure 8. Random Forest feature importance.

Topics (text-derived features). The topic features indicate that evasive action is most likely in operational contexts where visual acquisition is difficult, communication demands are high, or aircraft converge unexpectedly. The strongest predictor, Topic 3 (Visual Sightings and See-and-Avoid Encounters), reflects situations where pilots encounter helicopters, drones, gliders, or other traffic visually and often at close range; these encounters compress detection windows and frequently precipitate last-moment avoidance. Topic 2 (Tower Communications, Clearances, and Approach Sequencing) also ranks highly, capturing busy tower-controlled environments where sequencing, runway assignments, and rapid

exchanges can leave little margin during arrivals and departures. Topic 1 (Traffic-Pattern Operations and Training Flights) further highlights circumstances—often involving student pilots or dense pattern work—where spatial positioning and workload make timely detection and coordinated spacing more challenging. Topic 5 (Airspace Structure, Class Transitions, and Radar Contact Issues) contributes predictive value by signaling transitions into or out of controlled airspace, where radar identification, service changes, or class boundaries can yield uncertainty about relative positions. Topic 4 (Tower–Runway Interactions and Helicopter Operations) reflects runway-surface and helicopter-integration dynamics, which can create unexpected movement or wake interactions that tighten spacing. Collectively, the topic features delineate operational scenarios where visual, cognitive, or communication pressures increase the likelihood of evasive maneuvers.

Geometry and separation. Within these situations, geometric variables clarify how critical the encounter has become. Very small vertical or horizontal miss distances strongly distinguish events requiring corrective action from those resolved without maneuvering, capturing the physical proximity that most reliably predicts avoidance behavior. Relative distance in nautical miles plays a similar role: short slant range distances reduce recognition time, especially in environments where aircraft closure rates are high. These measurements enable the model's understanding of "too close for comfort" thresholds in ways that textual cues alone cannot.

Energy and phase-of-flight cues. Altitude above mean sea level (MSL) adds context about the flight phase in which the encounter occurs. Encounters near approach altitudes—where traffic density, configuration changes, and high radio activity converge—are more likely to require rapid conflict resolution. By contrast, en-route altitudes typically afford greater spacing and earlier detection, reducing the likelihood of evasive responses. Altitude, therefore, helps the model differentiate airport-area encounters from higher-airspace see-and-avoid scenarios.

Operational context. The 14 CFR Part under which Aircraft 1 is operating contributes structural information about expected procedures, equipment, and ATC integration. Part 121 and Part 135 flights generally operate under more standardized sequencing and communication protocols, whereas Part 91 flights exhibit more variability in traffic patterns, radio practice, and spacing expectations. These systematic differences help refine baseline probabilities of evasive action across diverse operational segments.

Collectively, the model highlights several practical points for improving safety performance. More specifically, strengthening radio practice during pattern operations and at tower runway interfaces can reduce ambiguity in dense traffic environments. Reinforcing expectations regarding pattern spacing and addressing base-to-final convergence hazards may help prevent many geometry-related conflicts before they develop. Increasing awareness of helicopter, glider, and UAS activity can mitigate the visibility and energy mismatches that often complicate timely acquisition. Prioritizing alerting and sequencing support during short slant range and low-altitude encounters can further reduce the likelihood of late detection. Finally, aligning procedures with the operating rules and equipment profiles ensures that guidance remains relevant and actionable. These measures directly target the conditions the model identifies as most predictive of evasive action and therefore represent meaningful opportunities for safety improvement.

Importantly, feature-importance scores reflect predictive association within the Random Forest model, not causal effects. Thus, high importance indicates that the feature helps the model discriminate between evasive and non-evasive outcomes in this corpus; it does not by itself prove a causal role in pilot decision-making. Confirming causality would require alternative study designs (e.g., FOQA/radar linkage with quasi-experimental methods or instrumental variable analysis), which we identify as recommended future work.

5. Discussion

The purpose of this study was threefold: to uncover latent themes in ASRS narratives using topic modeling, to develop an empirical typology of NMAC incidents through cluster analysis, and to predict the likelihood of pilot evasive action using supervised models that integrate narrative and structured features. The findings across these pillars point to a consistent conclusion: communication demands, encounter geometry, and operational context jointly shape both the emergence of NMAC risk and the decision to maneuver.

Topic modeling. The seven-topic solution provides an interpretable map of NMAC narratives that spans the major operational settings in which conflicts arise. These include traffic-pattern operations and training flights (Topic 1), tower communications, clearances, and approach sequencing (Topic 2), visual sightings and see-and-avoid encounters involving helicopters, drones, and gliders (Topic 3), tower–runway interactions and helicopter integration (Topic 4), airspace structure, class transitions, and radar contact issues (Topic 5), final-approach geometry and base-to-final convergence (Topic 6), and TCAS/ATC traffic advisories and altitude/heading management (Topic 7). Especially prominent are narratives involving compressed timing—short final, base-to-final turns, sequencing changes, and runway reassignment—where small lapses in spacing or delayed handoffs can rapidly narrow the window for see-and-avoid. Visual-encounter narratives involving gliders, helicopters, or drones highlight conspicuity and speed-mismatch challenges that reduce detection margins. In this way, the topics function as measurable constructs that translate narrative semantics into operational variables linked to workload, visibility, communication, and geometry.

Cluster analysis. A two-step clustering procedure yielded a four-cluster solution with strong separation, as supported by both the silhouette coefficient and the Calinski–Harabasz index. The clusters correspond to distinct operational segments shaped by phase of flight and operating variables. Topic-weight profiles clarify these distinctions: clusters characterized by tower sequencing and approach clearances load heavily on Topic 2, clusters dominated by VFR pattern work and training activity load on Topic 1, clusters with visual sightings of helicopters, drones, or gliders show strong Topic 3, and clusters representing en-route or transitional groups with traffic advisories load on Topic 7. Thus, the typology is not merely statistical—it aligns directly with the operational settings revealed by the topic model. It also clarifies where procedural structure is strongest (tower-directed approaches), where it thins (non-towered pattern work), and where mixed-equipage visibility mismatches are most salient.

Predictive modeling. Random Forest emerged as the champion model with consistent performance across partitions: accuracy in the high 0.87–0.88 range, misclassification around 0.12, specificity around 0.84–0.85, recall around 0.92, and F1 around 0.88. Discrimination metrics were strong (Gini $\approx$ 0.925, ROC AUC $\approx$ 0.962, lift $\approx$ 1.7), indicating that the model meaningfully improves on baseline classification. The feature-importance results reinforce the narrative patterns uncovered by topic modeling: Topic 3 (visual sightings), Topic 2 (tower sequencing), Topic 1 (pattern operations), Topic 5 (airspace transitions), and Topic 4 (runway-tower interactions) are among the strongest predictors. Geometry variables, including small vertical separation, horizontal miss distance, and short slant range, sharpen the decision boundary, while altitude helps distinguish airport-area encounters from en-route operations. 14 CFR Part and operation type shift the baseline by encoding procedural structure and equipment/crew differences. Thus, these signals show that evasive action is not random; it is the expected outcome when coordination load, visual acquisition challenges, and closure dynamics outpace shared situational awareness.

Integration across methods. The three analytical pillars converge. The topic model quantifies coordination load, visibility challenges, and geometry-related complexity. The

cluster analysis locates these conditions in distinct operational segments. The predictive model confirms that these same conditions significantly influence whether pilots maneuver. In short, the same narrative phrases pilots use to describe spacing, sequencing, airspace transitions, and traffic advisories are the semantic cues the model finds most informative, demonstrating that natural language is not noise—it is an operational signal.

Robustness and validity checks. Model metrics were stable across resamples. Topic selection balanced coherence, perplexity, and interpretability, supporting construct validity. ANOVA of topic weights showed significant differences across clusters, supporting the typology. Feature-importance profiles were consistent across partitions, and discrimination metrics remained high. Although ASRS is voluntary and not verifiable against radar or FOQA, triangulation across topic modeling, clustering, and prediction strengthens confidence that the patterns reflect underlying operational conditions rather than reporting artifacts.

Overall, the results shift the unit of analysis from isolated events to evolving conditions. Narrative-derived topics quantify coordination difficulty, clusters locate these conditions operationally, and predictive modeling clarifies when those conditions culminate in evasive action. More importantly, this integrated approach shows that natural language is not only descriptive but predictive and interpretable. This supports a move from retrospective narrative reading to prospective risk sensing in Safety Management Systems.

6. Conclusions

Theoretical contributions. This study makes three contributions. First, it validates narrative-derived topics as structured representations of coordination load, visibility constraints, airspace transitions, approach geometry, and advisory-driven interactions. The seven-topic solution aligns with known operational segments and demonstrates that narrative semantics can be measured and modeled rather than treated as anecdote. Second, it links these constructs to pilot behavior by predicting evasive action with high discrimination using integrated narrative and structured features, moving the analysis from risk states to the decision that prevents collision. Third, it develops a data-grounded typology of NMAC incidents that identifies where human-performance constraints are most binding.

Practical contributions. For practitioners, the results offer pathways for safety improvement. Training programs can rehearse pattern spacing, base-to-final transitions, and short-approach scenarios where closure dynamics are most acute. Controllers can emphasize clarity during approach sequencing and runway interactions. Airport managers and flying communities can strengthen CTAF practice with standardized call templates and spacing practices. Regulators can incorporate narrative-derived indicators into SMS systems to track communication load and geometry pressure, rather than only event counts. Avionics vendors can explore alerting thresholds that blend altitude bands, short slant range geometry, and narrative-linked context. Safety analysts can use topic trends to monitor rising coordination load before NMAC frequency increases.

Limitations. ASRS is voluntary and de-identified, which introduces selection bias and limits triangulation with radar or FOQA. Some structured fields are incomplete or inconsistent, especially 14 CFR Part entries in mixed operations. The topic, cluster, and prediction models are tuned to this corpus; semantic drift or procedural change may reduce performance without monitoring. Additionally, the U.S. operational context may limit generalization to other regulatory environments.

Temporal heterogeneity. The ASRS corpus spans 1988–May 2025, a period that encompasses major procedural and surveillance changes (e.g., TCAS adoption, UAS introduction). These temporal shifts can produce semantic drift in narratives and changes in reporting rates that affect model performance and topic prevalence. Our current study used ran-

dom splits and held-out validation to test generalization across the full period, but it did not perform explicit time-series cross-validation. We therefore treat temporal stability as an open limitation and recommend that future work perform time-based validation and post-deployment monitoring to detect drift.

Future research. Future work should pair ASRS with FOQA or surveillance data to calibrate geometry and timing, study temporal drift in topics and predictive features, compare LDA-based features with modern language-model embeddings, test the effects of procedural changes using quasi-experimental designs, incorporate UTM/Remote ID to examine mixed-equipage encounters, and evaluate whether topic-informed briefings or tuned alerting thresholds reduce real-world evasive maneuvers. Post-deployment monitoring should track the predictive model performance metrics (accuracy, misclassification, recall, F1, Gini, ROC AUC, and lift) to validate operational impact.

In conclusion, the results of this study demonstrate that thematic topics, geometry, and operational context can be integrated to anticipate when pilots will need to maneuver. Targeted improvements in radio practice, sequencing, and mixed-equipage integration can shift conditions away from last-second decisions. This approach offers a pathway from narrative reading to predictive sensing and supports safety interventions aligned with the conditions most likely to lead to evasive action.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in the NASA Aviation Safety Reporting System (ASRS) at https://asrs.arc.nasa.gov/search/database.html (accessed on 1 October 2025).

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

References

1. Federal Aviation Administration. *Aeronautical Information Manual: Chapter 7, Section 7—Near Mid-Air Collision Reporting*; U.S. Department of Transportation: Washington, DC, USA, 2025. Available online: https://www.faa.gov/air_traffic/publications/atpubs/aim_html/chap7_section_7.html (accessed on 1 October 2025).
2. National Transportation Safety Board. 2025 Potomac River Mid-Air collision: Preliminary Report (Report No. DCA25MA108). 2025. Available online: https://www.ntsb.gov/investigations/Documents/DCA25MA108%20Prelim.pdf (accessed on 1 October 2025).
3. Ortiz-Lytle, C. NTSB Concerned over 15,000 near-Misses Between Helicopters and Planes at a DC Airport in Just Three Years Before Crash That Killed 67. *New York Post*. 11 March 2025. Available online: https://nypost.com/2025/03/11/us-news/ntsb-concerned-over-15000-near-misses-between-helicopters-and-planes-a-dc-airport-in-just-3-years-before-crash-that-killed-67/ (accessed on 1 October 2025).
4. Arizona Pilots Association. Near Mid-Air Collisions (NMACs): September 2025 Report. 2025. Available online: https://azpilots.org/safety/51170-near-mid-air-collisions-nmac-s-september-2025 (accessed on 1 October 2025).
5. Brooker, P. Reducing mid-air collision risk in controlled airspace: Lessons from hazardous incidents. *Saf. Sci.* **2005**, *43*, 715–738. [CrossRef]
6. Kochenderfer, M.J.; Edwards, M.W.M.; Espindle, L.P.; Kuchar, J.K.; Griffith, J.D. Airspace Encounter Models for Estimating Collision Risk. *J. Guid. Control Dyn.* **2010**, *33*, 487–499. [CrossRef]
7. Kuchar, J.K.; Yang, L.C. A Review of Conflict Detection and Resolution Modeling Methods. *IEEE Trans. Intell. Transp. Syst.* **2000**, *1*, 179–189. [CrossRef]
8. Withrow, C.A.; Reveley, M.S. *Analysis of Aviation Safety Reporting System Incident Data Associated with the Technical Challenges of the Vehicle Systems Safety Technology Project*; NASA/TM-2014-217900; NASA Glenn Research Center: Cleveland, OH, USA, 2014.
9. Rose, R.L.; Puranik, T.G.; Mavris, D.N.; Rao, A.H. Application of Structural Topic Modeling to Aviation Safety Data. *Reliab. Eng. Syst. Saf.* **2022**, *224*, 108522. [CrossRef]
10. Tang, J.; Piera, M.; Nosedal, J. Analysis of induced traffic alert and collision avoidance system collisions in unsegregated airspace using a colored Petri net model. *Simulation* **2015**, *91*, 233–248.
11. Stroeve, S. What matters in the effectiveness of airborne collision avoidance systems? Monte Carlo simulation of uncertainties for TCAS II and ACAS Xa. *Aerospace* **2023**, *10*, 952. [CrossRef]

12. Figuet, B.; Monstein, R.; Waltert, M.; Morio, J. Data-driven mid-air collision risk modelling using extreme-value theory. *Aerosp. Sci. Technol.* **2023**, *142*, 108646. [CrossRef]
13. Haselein, B.Z.; da Silva, J.C.; Hooey, B.L. Multiple machine learning modeling on near mid-air collisions: An approach towards probabilistic reasoning. *Reliab. Eng. Syst. Saf.* **2023**, *239*, 109915.
14. Nordlund, P.-J.; Gustafsson, F. Probabilistic noncooperative near mid-air collision avoidance. *IEEE Trans. Aerosp. Electron. Syst.* **2011**, *47*, 1265–1276. [CrossRef]
15. Tuncal, A. Air traffic controllers' perspectives on unmanned aerial vehicles integration into non-segregated airspace. *J. Aviat.* **2024**, *8*, 153–165. [CrossRef]
16. Vera, D.; Pereira, Á.; Rodrigues, N.; Molina, J.; García, A.; Fernández-Caballero, A. Optical flow-based obstacle detection for mid-air collision avoidance. *Sensors* **2024**, *24*, 3016.
17. Lee, H.; Madar, S.; Sairam, S.; Puranik, T.G.; Payan, A.P.; Kirby, M.; Pinon, O.J.; Mavris, D.N. Critical Parameter Identification for Safety Events in Commercial Aviation Using Machine Learning. *Aerospace* **2020**, *7*, 73. [CrossRef]
18. Demir, G.; Moslem, S.; Duleba, S. Artificial intelligence in aviation safety: Systematic review and biometric analysis. *Int. J. Comput. Intell. Syst.* **2024**, *17*, 279. [CrossRef]
19. Berges, P.; Shivakumar, B.; Graziano, T.; Gerdes, R.; Celik, Z. On the feasibility of exploiting traffic collision avoidance system vulnerabilities. *arXiv* **2020**, arXiv:2006.14679. [CrossRef]
20. Zhou, D.; Zhuang, X.; Zuo, H.; Wang, H.; Yan, H. Deep learning-based approach for civil aircraft hazard identification and prediction. *IEEE Access* **2020**, *8*, 103665–103683. [CrossRef]
21. Vinogradov, E.; Kumar, A.; Minucci, F.; Pollin, S.; Natalizio, E. Remote ID for separation provision and multi-agent navigation. In Proceedings of the 2023 IEEE/AIAA Digital Avionics Systems Conference (DASC), Barcelona, Spain, 1–5 October 2023; pp. 1–10.
22. Yang, C.; Huang, C. Natural language processing in aviation safety: Systematic review of research and outlook into the future. *Aerospace* **2023**, *10*, 600. [CrossRef]
23. Paradis, C.; Hong, C.; Matthews, B.; Davies, M.D.; Hooey, B.L. Kaona: Deep searching and curating data from aviation safety reporting systems. In Proceedings of the AIAA SciTech 2025 Forum, Orlando, FL, USA, 6–10 January 2025; Available online: https://ntrs.nasa.gov/citations/20240015153 (accessed on 1 October 2025).
24. Odisho, E.V.; Truong, D.; Joslin, R.E. Applying machine learning to enhance runway safety through runway excursion risk mitigation. *J. Aerosp. Inf. Syst.* **2022**, *19*, 98–112. [CrossRef]
25. Truong, D. *Data Science and Machine Learning for Non-Programmers: Using SAS Enterprise Miner*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2024.
26. NASA Aviation Safety Reporting System (ASRS). About ASRS Data. Available online: https://asrs.arc.nasa.gov/search/dbol/aboutdata.html (accessed on 1 October 2025).
27. NASA Aviation Safety Reporting System (ASRS). ASRS Coding Taxonomy. Available online: https://asrs.arc.nasa.gov/docs/dbol/ASRS_CodingTaxonomy.pdf (accessed on 1 October 2025).
28. Nanyonga, A.; Joiner, K.; Turhan, U.; Wild, G. Semantic Topic Modeling of Aviation Safety Reports: A Comparative Analysis Using BERTopic and PLSA. *Aerospace* **2025**, *12*, 551. [CrossRef]
29. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *Adv. Neural Inf. Process. Syst.* **2002**, *14*, 601–608.
30. Blei, D.M. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
31. Pion-Tonachini, L.; Makeig, S.; Kreutz-Delgado, K. Crowd labeling latent Dirichlet allocation. *Knowl. Inf. Syst.* **2017**, *53*, 749–765. [CrossRef] [PubMed]
32. Kuhn, K.D. Using Structural Topic Modeling to Identify Latent Topics and Trends in Aviation Incident Reports. *Transp. Res. Part C Emerg. Technol.* **2018**, *87*, 105–122. [CrossRef]
33. Röder, M.; Both, A.; Hinneburg, A. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15), Shanghai, China, 2–6 February 2015*; Association for Computing Machinery: New York, NY, USA, 2015; pp. 399–408.
34. Hair, J.F., Jr.; Black, W.C.; Babin, B.J.; Anderson, R.E. *Multivariate Data Analysis*, 8th ed.; Cengage: Boston, MA, USA, 2019.

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.