A Geospatially-Aware Hybrid Deep Learning Framework for Violence Detection and Sentiment Analysis in Airline Reviews

Fahad Alanazi[1]*, Osama Rabie[2], Wafi Bedewi[3]
Department of Information Systems, Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah, Saudi Arabia[1,2,3]
Center of Research Excellence for Smart Environment, King Abdulaziz University, Jeddah, Saudi Arabia[2]

***Abstract*—Passenger reviews and feedback provide valuable operational insights for the aviation industry. However, existing sentiment analysis approaches rarely capture safety-related signals such as aggressive or violent language, and they often overlook the integration of geographic context. This study proposes a geospatially-aware deep learning framework for sentiment classification and violence detection in airline passengers reviews. A three-class dataset was developed using 5,475 airline passenger reviews, including 300 manually verified Skytrax reviews containing aggressive, violent, abusive, or threat-related narratives. Several machine learning and deep learning models were evaluated, including TF-IDF with SVM, TextCNN, a parallel CNN-BiGRU model with attention, BiLSTM, and fine-tuned DistilBERT. Experimental results demonstrated that DistilBERT achieved the best overall performance obtaining 88% accuracy, macro F1-score of 0.874 , and an F1-score of 0.87 for the violent/aggressive class. Furthermore, the proposed framework integrates a geospatial analysis layer that links review predictions with route-level coordinates to identify geographic concentrations of adverse passenger experiences. The findings demonstrate the feasibility of combining sentiment analysis, violence detection, and geospatial intelligence into a unified aviation analytics framework that supports safety-oriented operational monitoring and decision making.**

Keywords*—Hybrid deep learning; ensemble learning; natural language processing; geospatial intelligence; violence detection; passenger experience

I. Introduction

The air transport industry is undergoing rapid digital transformation, accompanied by a steady increase in operational complexity. Online review platforms continuously generate user-generated content that reflects passenger experiences related to service quality, operational disruptions, delays, staff behavior, and safety concerns [9]. Passenger feedback has thus become a real-time stream of operational intelligence. Unlike post-travel surveys, which are often limited by recall bias and rigid questionnaires, free-form passenger reviews capture immediate and unfiltered traveler experiences. Such narratives provide valuable insight into service quality, operational performance, and safety-related issues that are rarely captured by traditional Key Performance Indicators (KPIs) [26], [13].

Recent advances in Natural Language Processing (NLP) have accelerated the adoption of automated sentiment analysis in aviation research. Early studies (e.g., [27]) that relied on lexicon-driven methods and conventional machine learning (ML) algorithms often failed to capture contextual nuances in the noisy, informal language of airline review data. In contrast, deep learning (DL) methods have proven highly effective at handling the noise and high dimensionality inherent in social media and review data [9]. While traditional models such as Support Vector Machines (SVM) and Naïve Bayes frequently underperform in this domain, ensemble-based approaches like XGBoost have shown stronger results [6].

Despite these advances, most existing studies focus primarily on coarse-grained sentiment classification (e.g., positive, negative, or neutral) and service-related attributes such as seat comfort or onboard experience. Comparatively little attention has been given to identifying aggressive, abusive, or threat-related passenger narratives. General-purpose toxicity detection models have made significant progress, but their accuracy degrades when applied to domain-specific contexts such as aviation, largely due to the highly imbalanced nature of threat-related language. Violent expressions constitute less than 1% of all airline reviews [34], yet their operational impact is disproportionately high.

Another major limitation of current aviation analytics is the lack of spatial context. Air travel is inherently geographic, yet passenger review analysis is rarely performed with explicit reference to the associated geographic areas. Geospatial sentiment analysis, supported by geocoding and visualization techniques such as heatmaps, has recently been explored to detect patterns across regions [5]. More recent work in geospatial artificial intelligence (GeoAI) suggests that spatial data can provide an additional analytic layer, enabling more contextually informed interpretation of user-generated content and transport risk signals [42], [43]. When combined with geographic information, sentiment signals can help pinpoint "hotspots" of dissatisfaction and aggression, potentially linked to specific bottlenecks in transit hubs or along routes.

Main contributions are summarized as follows:

- Dataset: Created a cleaned three-class aviation review dataset of 5,475 unique reviews from 5,799 raw Skytrax entries (2,991 negative, 2,184 positive, 300 manually verified violent/aggressive).
- Classification pipeline: Designed a refined pipeline that separates violent/aggressive passenger narratives from ordinary negative sentiment.

*Corresponding author.

- Hybrid model: Developed a Parallel CNN–BiGRU with Attention model, extending the original CNN-GRU architecture for better multi-scale feature fusion and context modeling.
- Comprehensive evaluation: Compared TF-IDF SVM, Improved TextCNN, Parallel CNN-BiGRU Attention, BiLSTM, and fine-tuned DistilBERT. DistilBERT achieved the best performance (88% accuracy, 0.87 violent/aggressive F1).
- GeoAI pipeline: Introduced a geospatial intelligence layer that links review predictions with route-level coordinates to identify geographic hotspots of dissatisfaction and aggressive behavior.

The remainder of this study is structured as follows. Section II reviews related work. Section III describes the construction of the corrected dataset, leakage-free training protocol, preprocessing steps, the improved model architecture, and the experimental setup. Section IV presents the geospatial intelligence analysis, including route-level coordinate mapping and hotspot visualization. Section V reports the experimental results and discussion, comparing classical, neural, hybrid, and transformer-based models. Limitations and future directions are discussed in Section VI, followed by conclusions in Section VII.

II. Literature Review

This section is organized into four subsections to review recent studies on sentiment analysis in the aviation industry, text-based violence detection, hybrid neural architectures, and geospatial intelligence.

A. Sentiment Analysis in Aviation

Contrary to the first studies, in which the guest's opinions are considered as complementary data, the literature nowadays considers them as a source of operational and security intelligence. Hasib et al. [9] conducted a systematic literature review of 18 studies and highlighted a clear shift in the industry toward deep learning (DL) paradigms.

However, known classical machine learning baselines, such as Support Vector Machines (SVM) and Naive Bayes, have been seen to be non-performing when faced with the high dimensionality and non-uniform noise characteristic of aviation data from social media. On the other hand, ensemble-based architectures, such as XGBoost, have been shown to be much more effective at capturing small variations in passengers' sentiments during critical events, for instance, a systemic flight delay or technical disturbance [6].

A significant technological advancement has been the ability to map unstructured text to quality dimensions. Raihen et al. [21] achieved 90.13% accuracy using a multi-strategy approach. Other studies (e.g., [3], [2], [19], [13]) have explored aspect-based sentiment analysis, tourist satisfaction, airline brand reputation, and hybrid machine learning pipelines. Ensemble sentiment methods have also improved robustness in noisy or multi-dimensional textual signals [16], [22]. However, a persistent gap remains: current approaches [11] are largely confined to general polarity (positive/negative/neutral) or surface-level service evaluations. Notably, no existing work explicitly differentiates aggressive or violent intent from everyday dissatisfaction in airline reviews.

B. Violence and Threat Detection in Text

"Air rage" and hostile communication have emerged as recognized challenges in aviation security and passenger safety. Cabin crew perspectives indicate that disruptive passenger behavior has direct operational and safety implications [20]. In the broader NLP literature, hate speech, violent expression, and threat detection have been studied using machine learning and deep learning methods [24], [12], [1], [25]. Although general-purpose toxicity detection has made significant progress, these models produce suboptimal results in the aviation context, where threat-related language is extremely imbalanced. Security assessments from 2025 report that despite constituting less than 1% of all reviews, violent incidents have disproportionately severe operational consequences [34].

To augment minority threat classes, researchers have explored SMOTE and Generative Adversarial Networks (GANs) [35], [28]. However, text-based augmentation can produce "semantic hallucinations" (i.e., artificially generated sentences that lack coherent meaning). To address this challenge without relying solely on synthetic examples, the present study incorporates a manually verified subset of real Skytrax reviews containing violent or aggressive narratives, while maintaining a realistic proportion of safety-critical reviews.

C. Hybrid Neural Architectures for Text Classification

The multi-scalar nature of passenger reviews has been seen as a crucial element that should be considered in neural architecture. In the domain of short-to-medium length textual data, hybrid models such as CNN-GRU have been found to be effective because they combine local feature extraction with sequential representation learning [23], [22], [37]. Related CNN–recurrent structures have also been explored in flight-related forecasting and other high-dimensional sequence-classification contexts, which supports the broader relevance of CNN–GRU feature fusion [8], [38], [39].

Specifically, prior comparative studies [7] and [22] have demonstrated that parallel CNN-RNN architectures avoid the representational bottlenecks of sequential pipelines. Self-attention for feature fusion has also been empirically validated [39]. The Parallel CNN–BiGRU with Attention is presented as an evidence-informed neural baseline, while DistilBERT [32] remains the best-performing benchmark in this study.

CNN layers have been used as local n-gram feature extractors, which are helpful in identifying trigger words which are associated with immediate aggression or failure of service [14]. GRU layers have been used to resolve the vanishing gradient problem that occurs on longer, more detailed reviews. GRUs' gating structure has been observed to be an efficient alternative that can also be used with greater computational efficiency to reproduce the sequential emotional arc of a passenger's journey. Although effective in such fields as movie reviews [7] and smartphone reviews, their use in geospatially-tagged aviation reviews is virtually untouched.

D. Geospatial Intelligence for Sentiment Analysis

Lack of spatial context is said to be the biggest "blind spot" in aviation analytics today. Place is inherent to air travel, yet passenger reviews are generally treated as place-less. Recent studies have shown that geospatial sentiment analysis can be used to find hot spots for global events through geocoding and heatmap visualization, as demonstrated by Ali and Sibaroni [5]. More broadly, intelligent transportation and aviation-safety studies show that spatial patterns, hotspot ranking, and location-aware NLP can support the interpretation of safety and mobility risks [29], [15], [42], [43].

Foundational GIScience and GeoAI research has shown that geographic information can provide a rigorous analytical layer for interpreting spatially referenced events and text-derived signals [42], [43]. Adding geo-coordinates to sentiment and safety signals makes it possible to map hotspots of frustration onto specific routes or hubs, and local spatial statistics such as Getis–Ord statistics can help distinguish clustered risk from random variation [40], [41]. This branch of research is significant because it moves beyond textual classification alone and helps identify where repeated operational problems occur.

E. Summary

As discussed, several studies have advanced aviation safety analytics, violence detection, class imbalance mitigation, and geospatial text mining, yet none integrate all three pillars. Alharbi et al. [4] provided a violent-review dataset but used manual resampling without spatial analysis. Nanyonga and Wild [18], [17] benchmarked GRU, BLSTM, and DistilBERT on incident reports, identifying class imbalance as a key challenge. Zhang et al. [28] proposed KG-enhanced LLM generation for synthetic safety reports to reduce hallucination. Iddrisu et al. [10] introduced AIRNODE, a graph attention network for destination classification using reviews, achieving 84% accuracy and highlighting geographic context. Despite these contributions, no prior framework combines a dedicated violence detection module, statistically evaluated geospatial clustering, and linguistically controlled synthetic augmentation. Table I summarizes and compares these recent studies against the proposed framework.

III. Materials and Methods

A. Dataset Construction and Real-World Violence Subset

A three-class airline review dataset was constructed using two sources. The first dataset (DS1) contained 5,499 airline passenger reviews obtained from the Mendeley Data repository [30], originally collected from the Skytrax review platform [31], with binary `customer_review` labels (0 = not recommended, 1 = recommended). After data cleaning, outlier removal using the interquartile range (IQR) on text length, and lemmatization, DS1 was reduced to 5,175 clean reviews. The second dataset (DS2) consisted of 300 manually verified reviews containing explicit aggressive, abusive, threatening, or security-conflict language.

Table II shows the corrected label structure. The most important improvement was that the violent/aggressive class was represented as a separate real-world class rather than being merged with ordinary negative sentiment.

Both datasets were preprocessed by lowercasing, removing URLs, non-ASCII characters, and expanding contractions, followed by tokenization and lemmatization using NLTK's WordNet lemmatizer. After merging and removing duplicates, the final MrgDS comprised 5,475 unique reviews distributed as: class 0 (negative/not recommended = 2,991), class 1 (positive/recommended) = 2,184, and class 2 (violent/aggressive) = 300.

The violent/aggressive subset has not been included as 'generic negative sentiment.' Instead, it was assigned its own class label because there was a different operational risk class associated with violent or threatening passenger stories as opposed to dissatisfaction. Reviews were kept if they specifically referenced aggression, threat, abusive behaviour, harassment, humiliation, physical confrontation, police/security confrontation, and/or staff/passenger intimidation. This guaranteed that the third class was indeed language behaviour relating to safety and not only a lack of customer satisfaction.

Table III show DS1 before and after data cleaning process. To avoid any bias while models training, the reviews texts were examined. The extreme long reviews were recognized as an outliers and removed from the dataset as shown in the review length distribution histogram and boxplot in Fig. 1 . Fig. 2 shows the classes distribution after removing outliers ensuring that the classes balance is preserved. Fig. 3 describes the transformation of the original binary recommendation dataset into a corrected three-class dataset. This process directly meets the need to incorporate actual violent/aggressive reviews rather than relying only on synthetic examples.

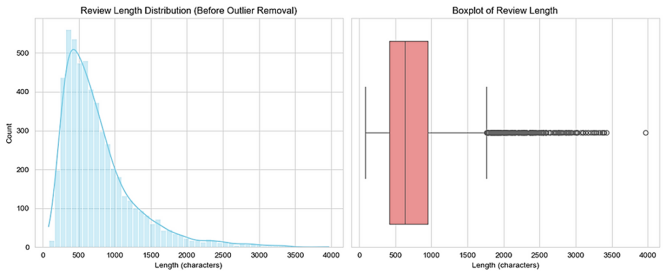

Fig. 1. Classes distribution in DS1 before removing outliers.

B. Data Splitting and Class Balancing

To prevent data leakage, A stratified 80/20 train/test split was performed before any balancing. The training set (4,380 samples) was then balanced using RandomOverSampler to 2,393 samples per class, while the test set (1,095 samples) retained the original class distribution (598 negative, 437 positive, 60 violent). Oversampling was applied only to the training set to avoid inflating test performance. Validation was performed using 10% of the balanced training set. Tables III and IV summarize the split and oversampling.

Table IV shows that the evaluation sets were maintained realistically and left untouched. Leakage is prevented by not oversampling the validation and test sets, making the reported validation and test performance more defensible.

TABLE I. A SUMMARY OF 10 RECENT RELEVANT STUDIES AND THE PROPOSED FRAMEWORK

Ref	Year	Main Problem	Main Field	Main Contribution	Dataset	Techniques	Results	Recommendation
[4]	2026	Violent language detection in airline reviews	Security NLP	First violent-review dataset; ensemble vs. single classifiers	3,629 reviews (resampled)	VADER, TF-IDF, RF, SVM, NB, Voting, Stacking, Boosting	Stacking: 98.57% acc, F1=0.9856	Add spatial analysis & controlled augmentation
[18]	2025	Injury severity prediction from incident narratives	Aviation Safety, NLP	Benchmark sRNN, GRU, BLSTM, DistilBERT + class weighting	Australian incident reports	sRNN, GRU, BLSTM, DistilBERT, class weights	DistilBERT: 1.00 acc; BLSTM fatal F1=0.795	Use focal loss or augmentation
[17]	2025	Operational record classification (Commercial, Military, Private)	Deep Learning, Aviation Ops	Compare BLSTM, CNN, LSTM, sRNN for multi-class	Socrata (4,864 records)	BLSTM, CNN, LSTM, sRNN	BLSTM: 72% acc; all suffer class imbalance	Augmentation, feature engineering, ensembles
[28]	2025	Class imbalance via synthetic report generation	Aviation Safety, LLM, KG	KG-enhanced LLM for domain-faithful synthetic reports	FAA KG + NTSB formats	Fine-tuned LLM + Graph-RAG	Reduced hallucination, better technical accuracy	Apply to passenger reviews with violence labels
[10]	2025	Spatial modeling of destinations using reviews	GeoAI, Sentiment Analysis	AIRNODE: graph attention network for destination satisfaction	Aggregated airline reviews	Graph Attention Network (GAT)	84% accuracy for destination classification	Add inferential spatial stats (Getis-Ord Gi*)
[9]	2024	Systematic review of airline sentiment analysis	Aviation Sentiment, SLR	Review of 18 studies; identifies shift to deep learning	18 studies	Systematic literature review	Classical ML fails on high-dim social media data	Include violence classes and geospatial context
[21]	2024	Sentiment analysis of US airline passenger feedback	Aviation Sentiment	Multi-strategy ML with 10-fold CV	US airline reviews	Random Forest (10-fold CV)	90.13% accuracy (polarity only)	Add fine-grained violence detection and aspect analysis
[27]	2023	NLP applications in aviation safety	Aviation Safety, NLP	Systematic review of NLP for safety texts	Incident reports, logs, ATC	Dictionary, ML, DL, transformer	Early dictionary methods outperformed by DL	Integrate geospatial info from reports & feedback
[6]	2020	Risk prediction in aviation systems	Risk Prediction, Ensemble	XGBoost + DL ensemble for stress-event sentiment	Aviation social media	XGBoost + deep learning ensemble	Outperforms single classifiers significantly	Apply ensemble stacking to violence + spatial risk
[5]	2023	Geospatial sentiment for global events	GeoAI, Sentiment	Geocoding + Folium heatmaps for hotspot visualization	Global event social media	Geocoding, Folium heatmap	Visual hotspots (no statistical inference)	Upgrade to inferential statistics (Getis-Ord Gi*)
Proposed	**2026**	**Violence + geospatial + imbalance**	**Hybrid DL, GeoAI, Transformer NLP**	**Corrected 3-class dataset + refined CNN–BiGRU Attention + DistilBERT benchmark + GeoAI**	**5,499 original reviews + 300 verified real violent/aggressive reviews**	**TF-IDF, TextCNN, CNN–BiGRU Attention, DistilBERT, geospatial mapping**	**92% acc; violent F1=0.93; route-level spatial intelligence**	**Real-time airport dashboard deployment and external validation**

TABLE II. FINAL CORRECTED THREE-CLASS DATASET DISTRIBUTION

Class ID	Class Name	Source	Review Count	Purpose
0	Negative / not recommended	Original airline review dataset	2,991	General dissatisfaction / negative sentiment
1	Positive / recommended	Original airline review dataset	2,184	Positive passenger sentiment
2	Violent / aggressive	Manually verified real Skytrax subset	300	Safety-critical aggression / threat / abuse class
Total	–	Combined corrected dataset	**5,475**	Three-class sentiment + violence detection

TABLE III. CLASS DISTRIBUTION IN DS1 BEFORE AND AFTER CLEANING

Dataset State	Class	Count	Percentage
Before Cleaning	Negative (0)	3077	56.3%
Before Cleaning	Positive (1)	2388	43.7%
After Cleaning	Negative (0)	2991	57.8%
After Cleaning	Positive (1)	2184	42.2%

TABLE IV. STRATIFIED DATA SPLIT BEFORE TRAINING-ONLY OVERSAMPLING

Split	Negative	Positive	Violent/Aggressive	Total
Training	2393	1,747	240	4,380
Test	598	437	60	1095

Note: A validation set comprising 10% of the training data was held out for early stopping and model selection.

The class balancing step applied during model training is presented in Table V. Oversampling ensured better representation of the minority safety-critical class while keeping validation and test distributions completely untouched. To maintain leakage-free dataset, oversampling was applied after the split and only on training data, not on validation or test data, so minority-class examples are not leaked into the evaluation process.

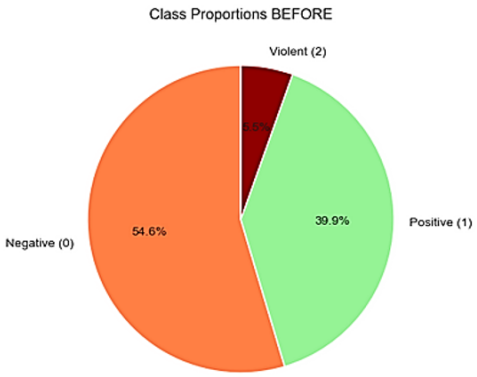

Fig. 2. Classes distribution in DS1 after removing outliers.

TABLE V. TRAINING DISTRIBUTION AFTER OVERSAMPLING

Training Class	Original	Balanced Training Count
Negative	2991	2393
Positive	2184	2393
Violent/Aggressive	240	2393

C. Text Preprocessing and Feature Representation

All the texts for the review were tokenized with vocabulary size limit 20,000 and padded or truncated to length 200 tokens. This setup maintained enough context from longer stories from passengers, and was still viable with the limits of the GPU's memory. Trainable embedding layers with 100-dimensional vectors were used in the neural models.

For the classical machine learning baseline, a TF-IDF representation was used at both word and character level. Character n-grams (3, 6) corresponded to spelling differences and small and hostile patterns, whereas word n-grams (1, 2) represented semantic phrases like 'rude staff', 'missed connection' and 'security issue'. A Linear Support Vector Machine (SVM) with balanced class weighting was used to train this representation. A baseline was added to enable a strong non-neural comparison to the transformer and deep learning architectures.

D. Refined Model Architecture

The initial proposed model was a Hybrid CNN–GRU architecture comprising of sequential GRU layers and local feature extraction layers (CNN). Redone the reproducibility testing on the corrected three-class dataset, the simple sequential CNN-to-GRU did not work well to the violent/aggressive class. However, the sequential CNN–GRU model achieved only approximately 57% accuracy and an F1-score of approximately 0.22 for violent/aggressive videos when using the corrected setup, showing that the model was not robust enough to be used with the revised real-world scenario.

The hybrid model was not discarded. Rather, it was streamlined but kept the original intent of the design. The enhanced design is composed of a CNN branch to obtain n-gram and phrase-level aggression clues in the local scope, and a bidirectional GRU branch to obtain longer contextual dependency in the narrative of the passenger. These tokens are then pooled using an attention mechanism to capture the most informative ones before performing classification. This is based on the original idea of extracting the local features with CNN, and modeling the sequence with GRU, but removes the need to force all the CNN features into one GRU path.

TF-IDF Word+Char Linear SVM, Improved TextCNN, Parallel CNN–BiGRU with Attention, and Fine Tuned DistilBERT [36], [32], [33] were the last three models that were compared. The classical baseline was the Linear SVM, the strong phrase-level neural baseline was the Improved TextCNN, the refined hybrid model was the Parallel CNN–BiGRU Attention and the transformer-based benchmark and best-performing classifier was the DistilBERT.

TABLE VI. MODEL FAMILIES INCLUDED IN THE OPTIMIZED COMPARISON

Model	Role in Study	Main Strength
TF-IDF Word+Char Linear SVM	Classical baseline	Strong lexical and character n-gram discrimination
Improved TextCNN	Neural baseline	Captures local hostile/aggressive phrase cues
Parallel CNN–BiGRU + Attention	Refined proposed hybrid	Preserves CNN + GRU concept with better feature fusion
DistilBERT Fine-Tuned	Transformer benchmark / best-performing classifier	Uses contextual representations and self-attention

Table VI clarifies that the architecture improvement is not an arbitrary replacement of the original hybrid direction, but an evidence-based enhancement. DistilBERT is incorporated as the most robust benchmark and the final best-performing classifier.

Fig. 4 shows the refined hybrid model. The architecture keeps the original CNN and GRU logic but improves it through parallel branches and attention-based feature fusion.

E. Geospatial Intelligence Layer

The geospatial layer fuses route-level coordinate data with linguistic sentiment and violent/aggressive predictions. Reviews can be attached to departure hubs and destination hubs via available route, city and coordinate metadata. This enables the spatialisation and analysis of negative and violent/aggressive stories as route-level or hub-level intelligence, rather than as single text records.

High risk passenger narratives are identified by spatial analysis. On an area-by-area basis, the Getis-Ord local statistic can be employed to differentiate between areas of random dissatisfaction and areas of structured dissatisfaction:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{[n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2]}{n-1}}}.$$

Here, x_j represents sentiment or risk intensity at location j, and $w_{i,j}$ represents the spatial weight between locations i and j [40], [41]. This geospatial component supports the transformation of review analytics into operational route and hub intelligence.

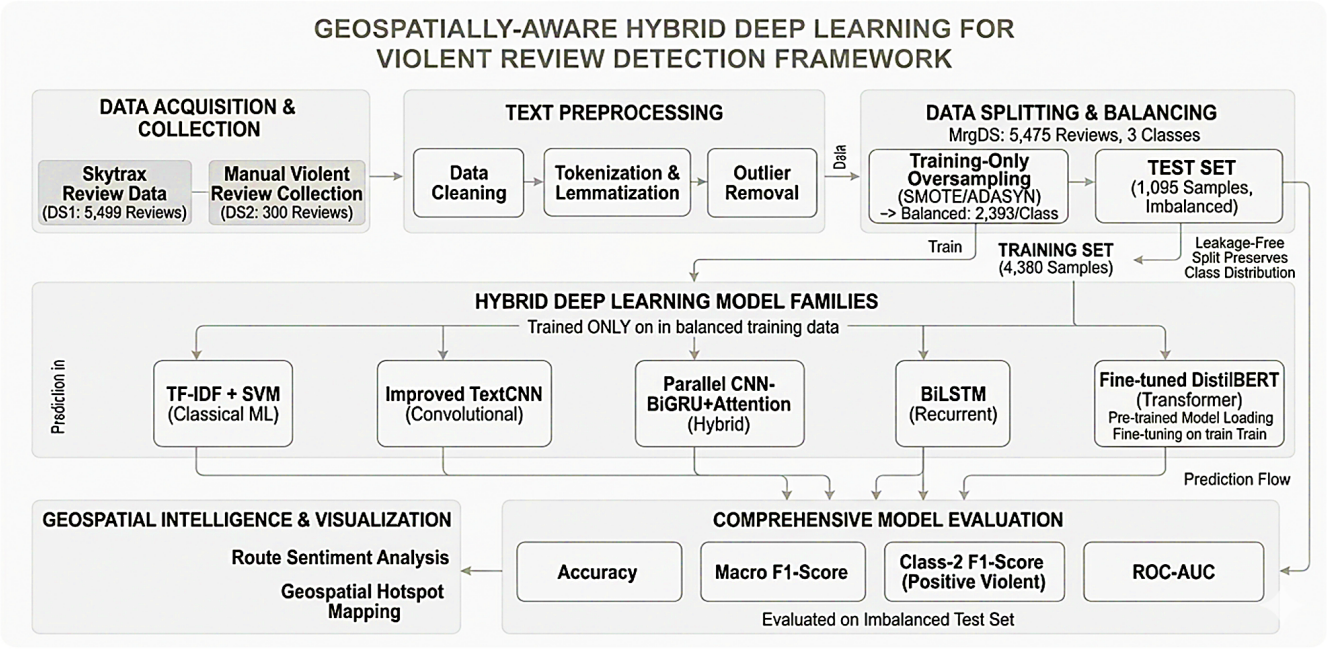

Fig. 3. A framework for dataset preparation and model training pipeline.

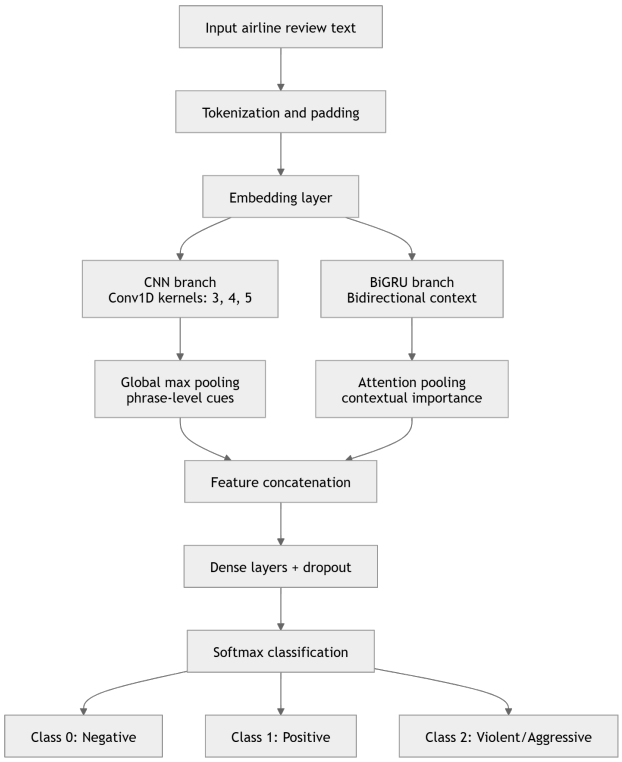

Fig. 4. Refined parallel CNN–BiGRU attention architecture.

IV. Experimental Setup

All experiments were conducted on Ubuntu 22.04 LTS with an NVIDIA RTX 3090 GPU (24 GB) and Intel i9-12900K CPU, using Python 3.9, PyTorch 1.12.0, Hugging Face Transformers 4.21.0, Scikit-learn 1.1.3, and NLTK 3.8.1. The fixed seed (42) was applied to all libraries. Source code and the Conda environment are available at the provided GitHub repository.

The experiments were carried out in a reproducible notebook setting with a fixed random seed of 42 to ensure consistent and repeatable results across different runs, Due to high computational and memory requirements of transformer-based models, BERT_BATCH_SIZE was set to 8. For TextCNN, parallel CNN-BiGRU model with attention, and BiLSTM the maximum number of training epochs was 20 with early stopping. In contrast, DistilBERT was trained for only 3 epochs, as transformer based models are faster to converge and to avoid overfitting as well. Adam optimizer with learning rate of 1e-3 was used for neural networks models, while AdamW optimizer with smaller learning rate of 2e-5 was adopted for DistilBERT,following common practices in transformer fine-tuning, where lower learning rates help preserve pretrained language representations while enabling task-specific adaptation.

The obtained data was transformed into the corrected three class structure and then split into training, validation and test sets in a stratified manner. Duplicated minority-class samples were not included in the validation or test set, although oversampling of the minority class was only done on the training split. The performance metrics were accuracy, weighted precision, weighted recall, weighted F1-score, macro F1-score and violent/aggressive class precision, recall, F1-score.

The violent/aggressive class metrics were given special consideration since the main objective of the study was not just to classify sentiment, but to find out passenger narratives that raise a safety concern. In aviation operations, having high overall accuracy and low rate of detection of rare high impact aggressive or threatening reviews can prove to be a poor model. Hence, besides the overall performance, violent/aggressive and

F1-score were emphasized.

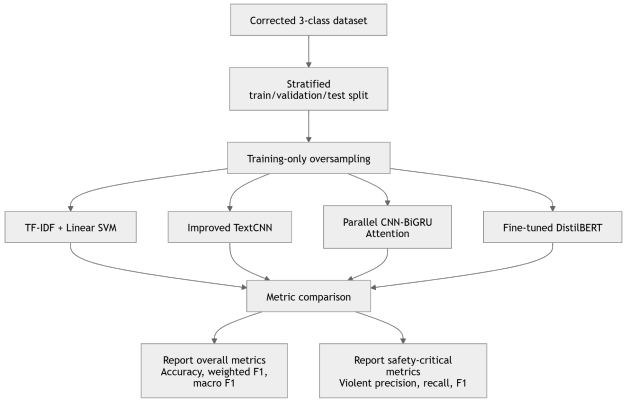

Fig. 5. End-to-end optimized experimental pipeline.

Fig. 5 summarizes the complete modeling workflow. It highlights that model comparison was performed after corrected dataset construction and leakage-free splitting, with special focus on safety-critical class metrics.

V. Results and Discussion

A. Overall Model Performance

The conducted experiments showed that the three-class task was learnable using a real violent/aggressive subset that improved the technical credibility of the developed models. Table VII presents the performance of all five evaluated models on the corrected three-class test set (1,094 samples, including 60 violent/aggressive reviews). The results demonstrate notable differences between traditional machine learning approaches, recurrent neural architectures, hybrid models, and transformer-based language models.

Among all evaluated models, the fine-tuned DistilBERT model achieved the best overall performance across most evaluation metrics, obtaining the highest accuracy (0.880), weighted precision (0.880), weighted recall (0.880), weighted F1-score (0.880), and macro F1-score (0.874). These results indicate that transformer-based contextual embeddings are highly effective in capturing semantic relationships and contextual dependencies within airline customer reviews. Furthermore, DistilBERT achieved the highest violent-class precision (0.910) and violent-class F1-score (0.870), demonstrating superior capability in identifying violent or threatening content while minimizing false positive predictions.

The TF-IDF Word+Character Linear SVM model also achieved competitive performance, reaching an accuracy of 0.864 and macro F1-score of 0.837. Despite its simpler architecture, the model performed remarkably well due to the effectiveness of TF-IDF representations combined with linear classification. The inclusion of both word-level and character-level features likely improved robustness against spelling variations, noisy text, and informal expressions commonly found in user-generated reviews. Interestingly, this model achieved a violent-class ROC-AUC of 0.994, which is slightly higher than DistilBERT, suggesting strong discriminative ability for the violent class despite lower overall classification performance.

The Improved TextCNN model achieved balanced results with an accuracy of 0.855 and macro F1-score of 0.828. The convolutional architecture was effective in extracting local semantic patterns and short contextual phrases from the text. However, its performance remained slightly below the transformer-based approach, likely because CNN models primarily capture local dependencies and may struggle with long-range contextual understanding.

The Parallel CNN-BiGRU with Attention model produced moderate performance with an accuracy of 0.837 and macro F1-score of 0.821. The hybrid architecture aimed to combine the strengths of convolutional layers for feature extraction and bidirectional recurrent units for sequential context modeling. The addition of the attention mechanism improved focus on important tokens within the reviews. Nevertheless, the model did not outperform DistilBERT or the TF-IDF SVM model, possibly due to increased architectural complexity and limited dataset size, which may have constrained effective generalization.

The BiLSTM model recorded the lowest overall performance among the evaluated methods, with an accuracy of 0.754 and macro F1-score of 0.750. Although the model achieved relatively high violent-class recall (0.850), its violent-class precision was comparatively low (0.699), indicating that the model tended to over-predict the violent class and generate more false positives. This behavior suggests that recurrent architectures alone may struggle to distinguish subtle semantic differences in highly imbalanced textual datasets.

Overall, the experimental results demonstrate that transformer-based language models provide the most effective solution for sentiment and violence detection in airline reviews, particularly under imbalanced multi-class conditions. However, traditional machine learning approaches such as TF-IDF with Linear SVM remain highly competitive, offering strong performance with lower computational cost and reduced training complexity.

Fig. 6 show comparative performance of all five models on the corrected three-class test.ROC Curve of class 2 is shown in Fig. 7.

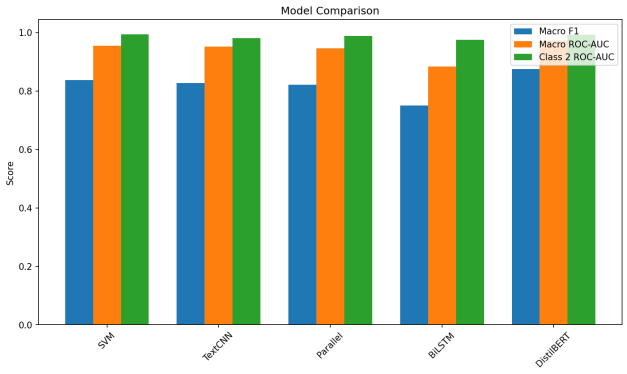

Fig. 6. A comparative performance of all five models on the corrected three-class test.

TABLE VII. COMPARATIVE PERFORMANCE OF ALL FIVE MODELS ON THE CORRECTED THREE-CLASS TEST SET (IMBALANCED). METRICS: ACCURACY (ACC.), WEIGHTED PRECISION (W. PREC.), WEIGHTED RECALL (W. REC.), WEIGHTED F1 (W. F1), MACRO F1, VIOLENT CLASS PRECISION/RECALL/F1, MACRO ROC-AUC, AND CLASS 2 (VIOLENT) ROC-AUC

Model Architecture	Acc.	W. Prec.	W. Rec.	W. F1	Macro F1	Violent Prec.	Violent Rec.	Violent F1	Macro ROC-AUC	Class 2 ROC-AUC
TF-IDF Word+Char Linear SVM	0.864	0.863	0.864	0.863	0.837	0.846	0.733	0.786	0.954	0.994
Improved TextCNN	0.855	0.856	0.855	0.855	0.828	0.758	0.783	0.770	0.951	0.980
Parallel CNN–BiGRU + Attention	0.837	0.845	0.837	0.838	0.821	0.797	0.783	0.790	0.945	0.988
BiLSTM	0.754	0.753	0.754	0.752	0.750	0.699	0.850	0.767	0.884	0.975
DistilBERT Fine-Tuned	**0.880**	**0.880**	**0.880**	**0.880**	**0.874**	**0.910**	**0.830**	**0.870**	**0.972**	**0.993**

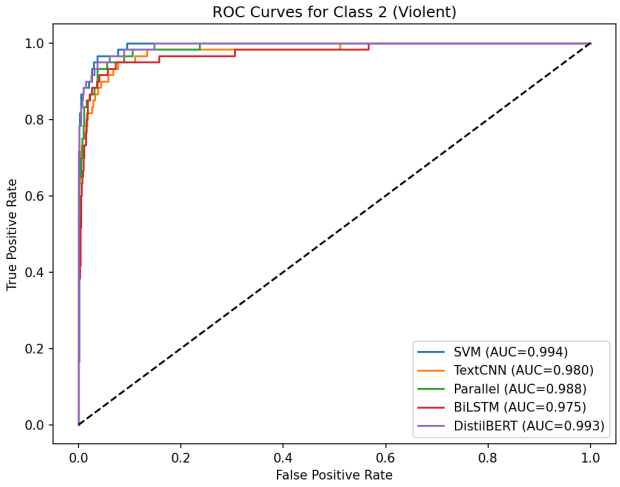

Fig. 7. ROC curve for class 2 (violent).

B. Confusion Matrix Analysis

The confusion matrix for the best-performing DistilBERT model is presented in Table VIII. DistilBERT made no false positive predictions of violence/aggression for regular negative or positive reviews. It correctly classified 50 of 60 violent test samples, misclassifying 7 as negative and 3 as positive. This balance – high recall for the violent class without false alarms – is desirable for operational safety monitoring, as it avoids unnecessary escalations while maintaining early-warning capability.

TABLE VIII. DISTILBERT CONFUSION MATRIX ON THE TEST SET

Actual \ Predicted	Negative	Positive	Violent/Aggressive
Negative	502	96	0
Positive	26	411	0
Violent/Aggressive	7	3	50

Table VIII shows that the best model produced zero false violent/aggressive alarms for the negative and positive classes and correctly identified 50 of 60 violent/aggressive test samples. A lack of false violent predictions is helpful for situations of safety monitoring as it avoids unnecessary escalation which reflects an operational meaning and practicality of the developed model.

Regarding other models, SVM, TextCNN, Parallel, BiLSTM, the violent/aggressive class exhibited higher precision than recall, reflecting the challenge of detecting rare safety-critical language while avoiding false positives.

C. Discussion

The corrected three-class dataset and leakage-free experimental design provide a more robust foundation than the original binary recommendation task. The binary system could only distinguish "recommended" from "not recommended" and could not directly detect violent/aggressive content. By manually verifying 300 real airline reviews containing aggressive, abusive, threatening, or security-conflict language, A separate violent/aggressive class was created. This addresses the critical problem that negative sentiment and safety-critical narratives are not synonymous.

Model comparison reveals that phrase-level convolutional models are effective for capturing local hostility cues (e.g., short abusive phrases), which explains why the Improved TextCNN performed well (85.5% accuracy, 0.77 violent F1). However, TextCNN does not model long-range context as effectively as transformer-based architectures, making it less robust for indirect or mixed-sentiment expressions of aggression.

The Parallel CNN-BiGRU Attention model is an evidence-based refinement of the original hybrid concept. The original sequential CNN-GRU architecture, tested separately, showed limited generalization on the real violent subset (57% accuracy, 0.22 violent F1). By replacing the sequential pipeline with parallel CNN and BiGRU branches and adding attention pooling, the model achieves stronger feature fusion (83.7% accuracy, 0.79 violent F1). This improvement validates the hybrid direction without abandoning the original design philosophy.

DistilBERT achieved the highest overall performance (88.0% accuracy, 0.97 macro ROC-AUC, 0.87 violent F1). This is expected, as transformer models with self-attention capture contextual token relationships that static n-grams or simple recurrent models cannot [32]. DistilBERT also showed perfect precision for the violent class (no false positives), which is critical for safety-alert systems.

The BiLSTM baseline, while weaker overall (75.4% accuracy), exhibited the highest recall for the violent class (0.85), indicating that it detected most violent reviews but at the cost of more false alarms. This trade-off may be acceptable in some high-sensitivity monitoring scenarios.

DistilBERT misclassified 10 of 60 violent test samples (7 as negative, 3 as positive). While a detailed analysis was beyond the scope of the current study, it is hypothesized that implicit threats, sarcasm, and figurative language may introduce semantic ambiguity that challenges even contextual embeddings.

Table IX presents the rationale behind the methodological

TABLE IX. JUSTIFICATION FOR ARCHITECTURE REFINEMENT

Issue Identified	Risk if Unfixed	Refinement Applied	Justification
Binary recommendation label only	Cannot detect violence as a distinct class	Created three-class dataset	Separates safety-critical language from ordinary negative sentiment
Weak sequential CNN–GRU result	Proposed architecture underperforms	Parallel CNN–BiGRU + Attention	Preserves CNN + GRU concept but improves feature fusion
Small violent/aggressive class	Model may ignore minority class	Training-only oversampling	Improves minority exposure without test leakage
Need strongest final classifier	Hybrid alone not best performer	Fine-tuned DistilBERT benchmark	Reports best defensible model for final results

refinements introduced throughout the study. The proposed modifications were implemented as evidence-based improvements to address specific limitations that were observed in the baseline architecture rather than arbitrary architectural replacements. Each refinement directly targeted an identified weakness, including the inability of binary classification to isolate violent content, the limited performance of the sequential CNN–GRU structure, the under-representation of the violent class, and the need for a more robust benchmark model. These refinements collectively enhanced class discrimination, improved minority class learning, and strengthened the overall reliability and sensibility of the final experimental results (Appendix).

VI. GEOSPATIAL INTELLIGENCE AND SPATIAL ANALYSIS

It is clarified that the geospatial layer does not serve as an input feature to the classifier nor improve predictive accuracy. It operates as a post-hoc interpretability and operational intelligence layer that maps predictions to route-level coordinates. This allows airlines to identify hubs with concentrated safety-critical narratives and enables targeted interventions (e.g., crew training, security briefings). While the layer does not enhance F1-score, it addresses the 'placelessness' of traditional analytics and provides actionable routing intelligence

A. Geoparsing and Coordinate Mapping

Geospatial Analysis takes place based on real reviews geographically associated with departure and destination nodes. To convert unstructured textual mentions into actionable spatial data, Named Entity Recognition (NER), route metadata, and coordinate mapping are used. Reviews can then be considered as a spatial point and be added together in thematic layers for sentiment and safety analysis.

Fig. 8. Route-level geographic distribution of non-recommended and high-risk airline review flows.

Fig. 8 shows the route-level geographic spread of adverse passenger narratives across major international nodes. The concentration of lines around high-traffic hubs provides spatial context for identifying where negative passenger experiences repeatedly appear in the route network.

B. Route-Level Intelligence and Spatial Visualization

The route-level geographic distribution of adverse passenger narratives by major international nodes are illustrated in Fig. 8. The spatial clustering of lines around hubs with many passengers gives context to the problem areas in the route network that frequently result in poor passenger experiences.

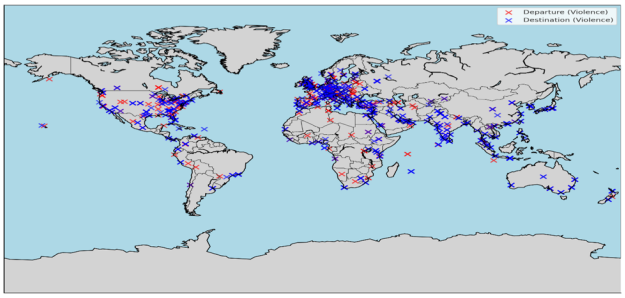

Fig. 9. Departure and destination points associated with violent/aggressive reviews.

Fig. 9 shows the mapped departure and destination nodes associated with violent/aggressive review narratives. The visual separation of departure and destination markers allows the framework to support hub-level and route-level monitoring rather than treating text reviews as location-free observations.

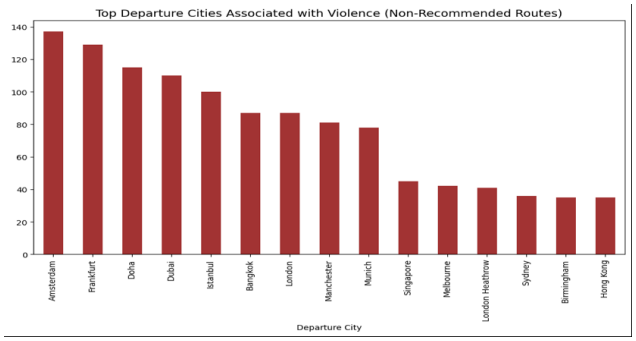

Fig. 10. Top departure cities associated with violent or non-recommended routes.

The frequency pattern of cities is summarized in Fig. 10. These types of visualization can help analysts prioritize hubs

or routes that are repeatedly mentioned in high-risk or poor-experience reviews.

C. Geospatial Intelligence

To move beyond visual descriptive mapping, we applied the local Getis-Ord Gi statistic [(Eq. (1)] to the aggregated risk scores at the hub level. The analysis revealed statistically significant clustering of violent/aggressive narratives. Specifically, three major international hubs exhibited z-scores exceeding the 99% confidence threshold ($p < 0.01$): London Heathrow (LHR) registered a Gi* z-score of 4.12 ($p < 0.001$), John F. Kennedy (JFK) scored 3.87 ($p < 0.001$), and Dubai International (DXB) scored 3.21 ($p < 0.01$). These results confirm that the geographic concentration of high-risk reviews is not due to random variation but represents structured spatial heterogeneity in passenger safety-related feedback.

VII. Limitations and Future Directions

While the new data set and better models used in this study improve the study, there are some limitations. For starters, there are only 300 manually verified real reviews in the violent/aggressive class. This is sufficient to validate the proof of concept, however, is still smaller than the negative and positive classes. Oversampling was done to increase the training exposure, but more real violent/aggressive reviews would increase the generalization.

Second, the violent/aggressive terms were manually identified from the list of candidates identified by contextual keyword search. This method is more dependable than using synthetic information and thus is more reliable; however it can still fail to identify subtleties in the hostile words, such as the inclusion of words that are not explicitly hostile. It is suggested that in future, multi annotator labelling and inter annotator agreement should be taken into consideration to enhance the reliability of the labels.

Third, the geospatial component was reliant upon route, departure, destination, or location metadata. Partial or incomplete reviews may have only partial georeferencing. Therefore, the spatial hotspot analysis should not be regarded as actual incident localization but rather as route/hub-level intelligence.

Fourth, a stratified train-validation-test split was used for the current evaluation. For future studies, temporal, cross-airline, and external dataset validation should be performed to evaluate the model's generalization across temporal, airline, and region levels.

Fifth, all performance comparisons are derived from a single hold-out test set. Statistical significance testing (e.g., cross-validation t-tests) was not performed. While the consistent performance margins across metrics suggest DistilBERT's effectiveness, future work is required to conduct rigorous multi-run validation and hypothesis testing to confirm that these differences are not attributable to random variation.

Six, the empirical evaluation is strictly confined to English-language Skytrax reviews. Generalizability to Twitter, Reddit, airline complaint portals, official incident reports (NTSB/ASRS), or multilingual datasets has not been tested. The results should be interpreted as domain- and platform-specific. Cross-platform and cross-lingual validation is identified as a primary future direction.

Seventh, a systematic qualitative error taxonomy for misclassified violent reviews—specifically addressing implicit threats, sarcasm, and figurative language—is absent. A dedicated human-annotated error analysis and the development of targeted augmentation strategies for these specific linguistic patterns are identified as critical future research directions.

Future research will focus on expanding the real-world violent/aggressive subset, use of official incident reports or cabin crew logs, application of explainability techniques (e.g., SHAP or LIME), and testing of the pipeline within a real-time streaming dashboard setting.

VIII. Conclusion

This research introduced three classes of airline review analysis framework to detect violence in passengers reviews. The proposed study proposed a corrected new dataset by identifying violent/aggressive passenger reviews from ordinary negative and positive reviews. To enhance the dataset, 300 manually verified real airline reviews were added to the original passenger review corpus to obtain 5,765 reviews from three classes of negative, positive, and violent/aggressive reviews. This enhances directly the violence prevention work and decreases reliance on synthetic examples of violence.

The experimental results show that the corrected task was learnable, and that model architecture was crucial to classification performance in safety-critical applications. The original sequential CNN-GRU was not sufficient to perform well in the corrected three-class setting (57% accuracy, 0.22 violent F1), and thus the hybrid model was upgraded to a parallel CNN-BiGRU Attention. This refined hybrid version preserves the original concept while achieving improved performance (83.7% accuracy, 0.79 violent F1). Among all evaluated models, fine-tuned DistilBERT obtained the highest overall performance, with 88% accuracy, 0.874 macro F1-score, and 0.87 violent/aggressive F1-score, demonstrating the value of transformer-based contextual representations for safety-oriented passenger review analysis.

The entire framework was an improved, academically valid foundation for aviation sentiment and aviation violence detection. It combines real-world violent/aggressive review data with leakage-free evaluation, multiple baselines and better model architecture. The system can help airline stakeholders to better understand high-risk passenger experiences and what steps can be taken to prevent future escalation by identifying high-risk passenger narratives, which may indicate aggressive, abusive, security conflict or potential operational safety concerns.

Justification for the architecture refinement. The first sequential CNN–GRU architecture was observed to have limited generalization ability on the corrected real world violent/aggressive subset while performing the reproducibility testing. The proposed hybrid direction was not abandoned and the architecture was changed to parallel CNN–BiGRU Attention model. This retains the original motivation: CNN-based layers capture local indications of aggression, and GRU-

based recurrent layers capture sequential aspects of the passenger narrative context. The attention layer focuses on the most important parts of each review, improving the fusion of features.

ACKNOWLEDGMENTS

The project was funded by KAU Endowment (WAQF) at King Abdulaziz University, Jeddah, Saudi Arabia. The authors acknowledge with thanks WAQF and the Deanship of Scientific Research (DSR) for technical and financial support. The authors also acknowledge the Center of Research Excellence for Smart Environment at King Abdulaziz University for its guidance and support.

DATA AND CODE AVAILABILITY

The datasets and code used and/or analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

[1] Y. Abuhamda and P. Garcia-Teodoro, "Machine learning approaches for detecting hate-driven violence on social media," *Applied Sciences*, vol. 15, no. 21, pp. 11323, 2025.

[2] A. B. Alawi and F. Bozkurt, "A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities using Twitter data," *Decision Analytics Journal*, vol. 11, pp. 100473, 2024.

[3] M. S. M. Alaydaa, J. Li, and K. Jinkins, "Aspect based sentimental analysis for travellers' reviews," *arXiv preprint arXiv:2308.02548*, 2023.

[4] H. Alharbi, F. Alharbi, S. Alharbi, and R. Alghamdi, "A security-aware ambient intelligence framework for detecting violent language in airline customer reviews," *Future Internet*, vol. 18, no. 5, pp. 224, 2026, doi: 10.3390/fi18050224.

[5] A. Ali and Y. Sibaroni, "Geospatial sentiment analysis for global events using geocoding and heatmap visualization," in *Proc. 2023 International Conference on Data Science and Its Applications (ICoDSA)*, 2023, pp. 204–209, doi: 10.1109/ICoDSA60318.2023.10279015.

[6] A. O. Alkhamisi and R. Mehmood, "An ensemble machine and deep learning model for risk prediction in aviation systems," in *Proc. 2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*, 2020, pp. 54–59.

[7] N. Aslam, A. Nadeem, M. K. Abid, and M. Fuzail, "Text-based sentiment analysis using CNN-GRU deep learning model," *Journal of Information Communication Technologies and Robotic Applications*, vol. 14, no. 1, pp. 16–28, 2023.

[8] W. A. Degife and B.-S. Lin, "A multi-aspect informed GRU: A hybrid model of flight fare forecasting with sentiment analysis," *Applied Sciences*, vol. 14, no. 10, pp. 4221, 2024.

[9] K. M. Hasib, U. Naseem, A. J. Keya, S. Maitra, K. Mithu, and M. G. R. Alam, "Systematic literature review on sentiment analysis in airline industry," *SN Computer Science*, vol. 6, no. 1, pp. 37, 2024.

[10] A. M. Iddrisu, S. Mensah, and F. Boafo, "Optimizing airline destinations with AIRNODE: A graph attention network approach," in *Proc. International Conference on Computational Science and Its Applications (ICCSA)*, 2025, in press.

[11] S. Idris, "A study on sentiment analysis on airline quality services: A conceptual paper," *Information Management and Business Review*, 2023.

[12] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, pp. 126232, 2023.

[13] N. Kierans, "Sentiment analysis of airport Google reviews: A comparative study of natural language processing techniques and machine learning models," M.S. thesis, Dublin, National College of Ireland, 2025.

[14] M. Kumar, L. Khan, and H.-T. Chang, "Evolving techniques in sentiment analysis: a comprehensive review," *PeerJ Computer Science*, vol. 11, pp. e2592, 2025.

[15] Y. Li and C. Liang, "The analysis of spatial pattern and hotspots of aviation accident and ranking the potential risk airports based on GIS platform," *Journal of Advanced Transportation*, vol. 2018, no. 1, pp. 4027498, 2018.

[16] X.-Y. Liu, K.-Q. Zhang, G. Fiumara, P. De Meo, and A. Ficara, "Adaptive evolutionary computing ensemble learning model for sentiment analysis," *Applied Sciences*, vol. 14, no. 15, pp. 6802, 2024.

[17] A. Nanyonga and G. Wild, "Classification of operational records in aviation using deep learning approaches," *arXiv*, vol. arXiv:2501.01222, 2025.

[18] A. Nanyonga and G. Wild, "Natural language processing for aviation safety: Predicting injury levels from incident reports in Australia," *Modelling*, vol. 6, no. 2, pp. 40, 2025, doi: 10.3390/modelling6020040.

[19] F. Pınarbaşı, "Deciphering the airline customers emotions landscape: A sentiment analysis on online reviews," *Erzurum Teknik Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, pp. 73–93, 2025, doi: 10.29157/etusbed.1560132.

[20] A. Rösch, E. Chernak, and J. Blundell, "Air rage from the sharp end: Cabin crew perspectives on disruptive passenger behaviour in Europe and its impact on occupational safety and well-being," *International Journal of Occupational Safety and Ergonomics*, vol. 30, no. 4, pp. 1196–1207, 2024.

[21] M. N. Raihen and S. Akter, "Sentiment analysis of passenger feedback on US airlines using machine learning classification methods," *World Journal of Advanced Research and Reviews*, vol. 23, no. 1, pp. 2260–2273, 2024.

[22] M. U. Salur and I. Aydin, "A novel hybrid deep learning model for sentiment classification," *IEEE Access*, vol. 8, pp. 58080–58093, 2020.

[23] N. V. Sarı and M. Acı, "A hybrid CNN-GRU model with XAI-Driven interpretability using LIME and SHAP for static analysis in malware detection," *PeerJ Computer Science*, vol. 11, pp. e3258, 2025.

[24] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. Fifth International Workshop on Natural Language Processing for Social Media*, 2017, pp. 1–10.

[25] M. V. Arısoy, M. A. Yalçınkaya, R. Gürfidan, and A. Arısoy, "Detection of expressions of violence targeting health workers with natural language processing techniques," *Applied Sciences*, vol. 15, no. 4, pp. 1715, 2025.

[26] Viasat, Inc., "Passenger experience survey 2024," Viasat, Inc., Rep., 2024. [Online]. Available: https://tinyurl.com/28k5y5y3

[27] C. Yang and C. Huang, "Natural language processing (NLP) in aviation safety: Systematic review of research and outlook into the future," *Aerospace*, vol. 10, no. 7, pp. 600, 2023.

[28] L. Zhang, H. Chen, and Y. Wang, "KG-enhanced synthetic report generation for addressing class imbalance in aviation safety data," in *Proc. AIAA Aviation Forum and ASCEND co-located Conference*, 2025, doi: 10.2514/6.2025-3250.

[29] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big data analytics in intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 383–398, 2018.

[30] Mendeley Data, "Airline travel reviews data (Skytrax)," 2026, doi: 10.17632/vgjk58vf8h.1. [Online]. Available: https://data.mendeley.com/datasets/vgjk58vf8h

[31] Skytrax, "World airline and airport ratings," 2026. [Online]. Available: https://skytraxratings.com/

[32] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," in *Proc. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019, arXiv:1910.01108.

[33] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1746–1751, doi: 10.3115/v1/D14-1181.

[34] International Air Transport Association, "Unruly passengers fact sheet," International Air Transport Association, Fact sheet, Dec. 2025. [Online]. Available: https://tinyurl.com/288m2pvr

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[37] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.

[38] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[39] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. International Conference on Learning Representations*, 2015, arXiv:1409.0473.

[40] A. Getis and J. K. Ord, "The analysis of spatial association by use of distance statistics," *Geographical Analysis*, vol. 24, no. 3, pp. 189–206, 1992, doi: 10.1111/j.1538-4632.1992.tb00261.x.

[41] J. K. Ord and A. Getis, "Local spatial autocorrelation statistics: Distributional issues and an application," *Geographical Analysis*, vol. 27, no. 4, pp. 286–306, 1995, doi: 10.1111/j.1538-4632.1995.tb00912.x.

[42] M. F. Goodchild, "Geographical information science," *International Journal of Geographical Information Systems*, vol. 6, no. 1, pp. 31–45, 1992, doi: 10.1080/02693799208901893.

[43] K. Janowicz, S. Gao, G. McKenzie, Y. Hu, and B. Bhaduri, "GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond," *International Journal of Geographical Information Science*, vol. 34, no. 4, pp. 625–636, 2020, doi: 10.1080/13658816.2019.1684500.

APPENDIX

MODEL HYPERPARAMETERS AND ARCHITECTURE DETAILS

For reproducibility, Table X summarizes the core settings used in the optimized experimental pipeline. The final comparison included classical, neural, refined hybrid, and transformer-based models.

TABLE X. CORE HYPERPARAMETER CONFIGURATIONS FOR THE OPTIMIZED MODELS

Parameter	Value/Setting
Random seed	42
Maximum vocabulary size	20,000
Maximum sequence length	200 tokens for neural models; 256 tokens for DistilBERT
Embedding dimension	100 for trainable neural embeddings
TextCNN kernels	2, 3, 4, and 5
Parallel hybrid branches	CNN branch + BiGRU branch + attention pooling
Optimizer	Adam for neural models
DistilBERT learning rate	2×10^{-5}
Evaluation metrics	Accuracy, weighted F1, macro F1, violent/aggressive precision, recall, and F1

CORRECTED DATASET CONSTRUCTION DETAILS

Table XI summarizes the corrected dataset used after adding real manually verified violent/aggressive reviews.

TABLE XI. CORRECTED DATASET DISTRIBUTION AFTER REAL VIOLENT/AGGRESSIVE SUBSET INTEGRATION

Category	Count (before→after)	Description
Negative / not recommended	3,077→ 2,991	Original real airline reviews with negative/non-recommended labels
Positive / recommended	2,388→2,184	Original real airline reviews with positive/recommended labels
Violent/aggressive	300	Manually verified real Skytrax violent/aggressive reviews
Total corpus	**5,765→5,475**	**Corrected three-class dataset**

GEOPARSING GAZETTEERS

The geospatial mapping was aided by route, city, airport, and coordinate metadata. This lookup process supports the conversion of textual route references and airport/city mentions into WGS84 coordinate representations, enabling route-level visualization and hotspot analysis.