Optimizing Airline Destinations with AIRNODE: A Graph Attention Network Approach

Abdul Raheem Shaik[1], Abdulaziz Alhamadani[2], Shailik Sarkar[3], and Chang-Tien Lu[3]

[1] Bradley Department of Electrical and Computer Engineering, Virginia Tech, Falls Church, VA 22043 USA
ashaik97@vt.edu

[2] Department of Data Science and Business Analytics, Florida Polytechnic University, Lakeland, FL 33805 USA
aalhamadani@floridapoly.edu

[3] Department of Computer Science, Virginia Tech, Falls Church, VA 22043 USA
{shailik, ctlu}@vt.edu

Abstract. The airline industry faces the critical challenge of meeting increasing passenger expectations amidst rapid technological advancements and intense competition. To remain competitive, airlines must gain a deeper understanding of passenger satisfaction and use this knowledge to improve service quality. This paper addresses this challenge by leveraging online customer reviews to derive actionable insights into passenger experiences. We introduce AIRNODE (Attention-based Insights for Reviewing Node Optimized Destinations and Experiences), a comprehensive two-stage model designed to analyze these reviews. AIRNODE constructs a weighted graph to aggregate review data and utilizes a Graph Attention Network (GAT) to model complex spatial relationships between destinations, achieving 84% accuracy in classifying destinations based on aggregated user satisfaction. Through advanced keyword extraction, we identify key aspects such as customer service, delays, and staff behavior, providing deep insights into the factors that drive passenger satisfaction. Case studies highlight destinations with varying levels of satisfaction, identifying positive attributes and areas needing improvement, and offering detailed insights and justifications for enhancing customer satisfaction. These insights equip airlines with a data-driven strategy to enhance service quality, meet traveler expectations, and maintain a competitive edge in a dynamic market.

Keywords: Airline Industry · Large Language Models· Graph Neural Networks· Online Customer Reviews Analysis

1 Introduction

The travel and airline industries are undergoing rapid transformations due to advancements in technology, changes in consumer behavior, and increased competition. These dynamics make it essential to understand and improve passenger experiences to remain competitive. Examining the quality of passenger experiences and exploring the potential of utilizing online customer reviews to enhance service offerings is crucial for the field. By leveraging insights from customer feedback, airlines can identify key areas for improvement and develop strategies to enhance overall service quality.

Understanding what drives passenger satisfaction at various travel destinations is crucial for the industry. Identifying and analyzing the factors that contribute to positive and negative experiences will reveal what makes a destination highly satisfying or particularly unsatisfactory for travelers. Gaining such insights is vital for improving service quality and passenger satisfaction. Prior studies have demonstrated that multiple factors, including service quality, destination attractiveness, and overall travel convenience, significantly influence passenger satisfaction [6,2]. By understanding these elements, service providers can prioritize improvements and create more satisfactory travel experiences, leading to increased customer loyalty and a competitive edge in the market.

Additionally, the fierce competition in the airline industry necessitates that airlines continuously enhance customer satisfaction by understanding passenger expectations and refining service offerings. The surge in social media usage has produced an abundance of online customer reviews, offering airlines a rich source of insights into their services and those of their competitors. Harnessing advanced text analysis techniques, especially large language models (LLMs), can efficiently extract actionable intelligence from these reviews. This allows airlines to gain critical insights that drive service enhancements and inform competitive strategies [14,5].

Existing works in online review analysis have used various NLP techniques and deep learning models. Some of the works have used topic modeling to categorize not only the important topics but also words relevant to each topic for airline companies to identify major service failures where further incorporation of sentiment analysis has contributed to a more accurate identification of negative experiences [13,4]. However, the statistical nature of topic modeling fails to capture the contextual length of a text that deep learning models can capture. Deep learning techniques have largely been used in this domain as a tool for individual classification of online reviews [22,20] and analysis of common themes associated with negative experiences through sentiment analysis [3,17]. However, these works do not address the issue of explainability in traditional deep-learning models. The lack of interpretability is a problem faced by many in this domain, as meaningful insights behind a predictive algorithm are a necessity for the successful deployment of such systems in the real world. Furthermore, there is no significant work in analyzing the spatial characteristics of online text reviews related to airports as the geographical location and characteristics of these airports will play a role in shaping the expectation or requirement for an ideal customer experience.

Hence, we aim to address the following research questions:

- **RQ1:** *How can we model the textual information of online reviews for each airport instead of treating each review as an individual datapoint?*

- **RQ2:** *How can the spatial correlation and dependency between different destinations be modeled?*
- **RQ3:** *How do we provide feedback or explanation for the positive or negative trend in online reviews for each destination classification?*

In this paper, we introduce the two-stage model AIRNODE (Attention-based Insights for Reviewing Node Optimized Destinations and Experiences) to address the challenges of destination recommendation in the airline industry. To answer RQ1, we construct a weighted graph where nodes represent destinations, enriched with RoBERTa word embeddings and sentiment analysis scores to effectively model the aggregated textual information of online reviews. For RQ2, we utilize a Graph Attention Network (GAT) that captures complex inter-destination relationships through dynamic attention mechanisms, allowing for the modeling of spatial correlations and dependencies between destinations. Addressing RQ3, we extract keywords from customer reviews and engineer an advanced Large Language Model (LLM) setup to analyze these keywords. This setup uses case-specific designed prompts to identify both positive attributes that enhance visitor satisfaction and areas needing improvement based on customer concerns, offering detailed insights and justifications for these findings. The integration of GAT and LLMs, driven by these targeted prompts, offers an efficient framework for extracting meaningful insights from unstructured data, providing a data-driven strategy to enhance customer satisfaction. By leveraging these insights, airlines can improve customer experiences, customize their service offerings to meet traveler expectations, and increase customer loyalty and competitiveness in the market. Our contributions are summarized as follows:

- Developed AIRNODE, a two-stage framework for destination recommendation, integrating a weighted graph representation enriched with RoBERTa embeddings and sentiment analysis scores.
- Formulated a case-specific Graph Attention Network architecture to capture complex inter-destination relationships and leverage unstructured customer review data for destination node classification based on customer reviews and satisfaction metrics.
- Engineered an advanced Large Language Model setup with specific prompts to identify positive attributes and areas needing improvement, providing detailed insights and justifications for enhancing customer satisfaction.

2 Related Work

2.1 Text classification of Online Reviews

Text classification of customer reviews has been an important application for several industries. Recent works have addressed several challenges and proposed novel approaches like Clue And Reasoning Prompting (CARP) for domain-specific text classification tasks[23]. In recent work, [26] proposes RGPT, an adaptive boosting framework, to employ an ensembling technique to produce specialized text classification models producing performance comparable to state-of-the-art language models. Moreover, [17] explored how sentiment analysis and customer rating prediction can improve the accuracy of a recommendation system showing significant improvements in prediction metrics. Similarly, [21] focuses on this task by leveraging the associated topics and attributes. The importance of both lexicon-based approach and deep learning models was demonstrated by [3] where they analyzed how combining TextBlob with different machine learning and deep learning models can improve classification accuracy. These works have showcased the effectiveness of machine learning models in text classification for customer reviews. However, they focus solely on text attributes and don't consider the interaction space of online review platforms. To this effect, [16] proposes the framework "review network feedback" to leverage interpersonal interactions among customers, resulting in superior performance over existing baselines in online review datasets. While the existing works contribute significant insights into text classification tasks for online reviews, our work differs from this approach by deviating from the black-box nature of predictive modeling and instead allows for more interpretable results.

2.2 Customer Feedback Analysis in the Aviation Industry

Analysis of online customer reviews is seen as an effective tool for identifying factors affecting customer experience and consequently improving services to facilitate informed decision-making. Different NLP and data mining techniques have been utilized in this regard including but not limited to topic modeling, sentiment analysis, and deep learning-based text classification. With the popularity of language models and large language models, this field has seen more and more applications of AI tools for such analysis. The potential for embeddings created by language models has been demonstrated by works that have extensively used pre-trained models like BERT, a popular language model[12]. The utilization of a fine-tuned large-language model has emerged to be particularly effective for domain-specific tasks[23]. Sentiment analysis and topic modeling have also been widely applied to understand passengers' opinions and experiences, where different methods like VADER, LDA or logistic regression have been explored for domain-specific tasks like identifying contributing topics and words to particularly negative experiences[10,13,21]. These strategies have proved to be effective in identifying key issues and sentiment drivers in airline reviews, revealing the importance of factors such as seat, service, meal, and delays[13].

The challenges and nuances of analyzing aviation-related customer reviews are manifold and existing literature has addressed this issue through the implementation of various advanced techniques like sarcasm detection[11], deep learning algorithms[8], multimodal approaches[24], time series methods[25], and aspect-based sentiment analysis[1]. The capability of these approaches to capture sudden changes in passenger sentiments can potentially help airlines take appropriate mitigatory measures. Moreover, it is well demonstrated how the ubiquity of accessible open-source data collected from blogs and online customer review platforms can be utilized by text-extracting software to assess the level of services perceived by airport customers[7]. While the existing literature has made significant contributions to the field, they almost exclusively focus on individual reviews and are mostly interested in airline-specific analysis of customer reviews. However, in this paper, we hypothesize that the airports and the facilities available in those specific locations also shape the customer experience. Therefore, it becomes essential to model the spatial correlation between different airports and analyze the customer reviews from that perspective. Furthermore, there also exists a gap in this field regarding utilizing the capabilities of LLM to provide interpretable explanations to human actors in decision-making. Our method addresses these limitations by utilizing Graph Neural Networks for spatial modeling of aggregated texts while also leveraging unsupervised learning techniques and LLMs to generate explanations for each prediction. This multi-faceted approach in the analysis of customer feedback in the aviation industry has the potential to create a novel framework for more informed decision-making and improve the quality of service.

3 Problem Statement

3.1 Introduction and Definition of the Problem

The advent of big data analytics in the airline industry offers unparalleled opportunities to enhance customer experiences by analyzing extensive volumes of passenger feedback. This study aims to harness this wealth of data through a sophisticated graph-based model that classifies destinations based on aggregated user satisfaction. This classification is important because it allows airlines to identify which destinations are meeting customer expectations and which ones need improvement. Destinations are represented as nodes, and journeys between them are depicted as edges in a graph structure. To effectively perform this classification, we leverage Graph Attention Networks (GAT).

3.2 Graph Construction and Feature Representation

The dataset is represented as a directed graph $G = (V, E)$, where each node $v_i \in V$ corresponds to a unique destination code from the dataset. Each edge $(v_i, v_j) \in E$ represents a journey between destinations, weighted by the travel distances. The feature vector for each node v_i incorporates multiple data sources. Textual embeddings are generated from reviews using advanced NLP models such as RoBERTa and BERT. Sentiment scores are extracted from the reviews using sentiment analysis techniques. Additional features include aggregated ratings like Entertainment, Food, Ground Service, Seat Comfort, Service, and Wifi Ratings. Formally, the feature vector for node v_i is expressed as:

$$\mathbf{x}_i = [\mathrm{emb}_i, s_i, r_i]$$

where emb_i represents the textual embeddings, s_i the sentiment scores, and r_i the aggregated ratings.

Edges between nodes are weighted based on normalized distances to reflect the travel distance between destinations:

$$w_{ij} = \frac{\text{Distance}(v_i, v_j)}{\max(\text{Distances})}$$

The adjacency matrix A is constructed such that $A_{ij} = w_{ij}$ if there is a journey between v_i and v_j, and $A_{ij} = 0$ otherwise.

3.3 Challenges

This study addresses several key challenges. First, handling sparse and imbalanced data is crucial due to the variability in the number of reviews per destination, which can lead to biased classification. Second, integrating multimodal data, which combines textual and numerical data into a coherent graph structure, is complex. Third, effective graph learning utilizing GAT to capture complex relationships within the graph for accurate classification is necessary.

4 Methodology

4.1 Overview of the Proposed Model

The methodology involves constructing a Graph Attention Network (GAT) model to analyze and classify destinations based on airline reviews. GAT is selected for its capability to dynamically weight the importance of neighboring nodes, thereby enhancing predictive accuracy in graph-based data.

4.2 Graph Attention Network (GAT)

The GAT model incorporates an attention mechanism to dynamically adjust the importance of each node's neighbors. The attention coefficients α_{ij} are computed as:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^\top[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_j]))}{\sum_{k \in \mathcal{N}(i)} \exp(\text{LeakyReLU}(\mathbf{a}^\top[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_k]))}$$

where $\mathbf{h}_i$ and $\mathbf{h}_j$ are the feature vectors of nodes i and j, $\mathbf{W}$ is a learnable weight matrix, and $\mathbf{a}$ is the attention mechanism's parameter vector. This mechanism allows the model to focus on more relevant nodes, thereby improving the learning process.

The GAT layer output is computed by aggregating the features of neighboring nodes, weighted by the attention coefficients:

$$\mathbf{z}_i = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}\mathbf{h}_j\right)$$

where σ is a non-linear activation function, LeakyReLU in this case. The attention coefficients α_{ij} ensure that the model focuses more on the relevant neighboring nodes during the aggregation process.

The architecture of the GAT model involves multiple GATConv layers to learn hierarchical feature representations. The model consists of the following layers:

1. **First GAT Layer**:

$$\mathbf{H}^{(1)} = \text{GATConv}(\mathbf{X}, \mathbf{A})$$

where $\mathbf{X}$ is the node feature matrix and $\mathbf{A}$ is the adjacency matrix.

2. **Dropout Layer** to prevent overfitting.
3. **Second GAT Layer**:

$$\mathbf{H}^{(2)} = \text{GATConv}(\mathbf{H}^{(1)}, \mathbf{A})$$

4. **Final GAT Layer**:

$$\mathbf{Z} = \text{GATConv}(\mathbf{H}^{(2)}, \mathbf{A})$$

This multilayer architecture allows the GAT model to capture complex relationships within the graph, enhancing its ability to accurately classify destinations based on satisfaction.

The model employs the Sparse Categorical Crossentropy loss function for binary classification:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability.

The model was optimized using the Adam optimizer with a learning rate scheduler:

$$\eta_t = \eta_0 \cdot \exp\left(-\frac{t}{\tau}\right)$$

where η_t is the learning rate at step t, η_0 is the initial learning rate, and τ is the decay step.

4.3 Keyword Extraction and Clustering Analysis

After training the GAT model, we used YAKE (Yet Another Keyword Extractor) to efficiently extract keywords from each review. YAKE was chosen for its speed and ease of implementation, making it suitable for large-scale datasets[19]. The extracted keywords help identify the main points discussed in the reviews.

To organize the keywords and sentiments into meaningful groups, we applied the K-means clustering algorithm, selected for its simplicity and effectiveness. The clustering process involved the following steps:

1. Combine extracted keywords, sentiment scores, and overall ratings for each journey. Formally, let $\mathbf{k}_i$ represent the set of keywords for review i, s_i the sentiment score, and o_i the overall rating. The feature vector for clustering is:

$$\mathbf{v}_i = [\mathbf{k}_i, s_i, o_i]$$

2. Apply the K-means clustering algorithm to segment the data into k clusters. K-means works by partitioning the dataset into clusters where each cluster is represented by its centroid. The algorithm aims to minimize the within-cluster sum of squares (WCSS):

$$\text{WCSS} = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

where C_j is the set of points in cluster j, x_i is a data point, and μ_j is the centroid of cluster j.

3. Calculate the frequency of keywords within each cluster. For each cluster C_j, determine the frequency f_{k_j} of each keyword k within the cluster:

$$f_{k_j} = \frac{\text{Number of occurrences of keyword } k \text{ in cluster } C_j}{\text{Total number of keywords in cluster } C_j}$$

The resulting clusters are analyzed to determine the main talking points, positives, and negatives using our engineered LLM setup.

4.4 Engineered Large Language Model

After clustering, we use GPT-4o to analyze the clusters. GPT-4o, known for its advanced natural language understanding and generation capabilities, was chosen for its superior performance, efficiency, and speed in handling complex language tasks and providing detailed, contextually accurate insights. Comparative analyses, such as the one detailed [9], highlight GPT-4o's advantages over other models, such as Claude 3, making it an ideal choice for our application.

The prompts used for this analysis were crafted to extract nuanced information from the clustered reviews. These prompts guide GPT-4o to identify positive attributes that should be maintained for visitor satisfaction and areas needing improvement based on customer concerns. By directing GPT-4o to focus on specific aspects of customer feedback, we ensure that the analysis is both comprehensive and actionable.

The prompts were derived based on the need to generate actionable insights from unstructured review data. They were crafted to direct the LLM to focus on specific aspects of customer feedback:

- Positive Aspects: Identifying and justifying attributes that should be maintained to ensure high customer satisfaction.
- Negative Aspects: Highlighting areas that require improvement and justifying how to address these concerns.

The insights generated are then compiled into reports and expressed in tabular format for clarity and ease of interpretation. By leveraging these insights, airlines can focus on maintaining strengths and addressing weaknesses in their services, ultimately improving customer satisfaction and loyalty. The following are example prompts used in the analysis:

- Positive Keywords Analysis:
  As an expert aviation service manager, analyze the provided list of positive keywords from airline journey customer reviews. Rank the keywords by frequency and identify which aspects should be maintained to satisfy visitors and increase travel to your airport. Include reasoning for your suggestions based on the keywords.
- Negative Keywords Analysis:
  As an expert aviation service manager, analyze the provided list of negative keywords from airline journey customer reviews. Rank the keywords by frequency and identify which areas need improvement to enhance visitor satisfaction and increase travel to your airport. Include reasoning for your suggestions based on the keywords.

The AIRNODE framework, summarized in Figure 1, illustrates our model which utilizes GAT to classify destinations based on aggregated user satisfaction. The figure depicts the process from data input and sentiment analysis, through GAT construction and classification, to keyword extraction and analysis using the engineered LLM setup with case-specific prompts to gain insights from review keywords on a destination-by-destination basis.

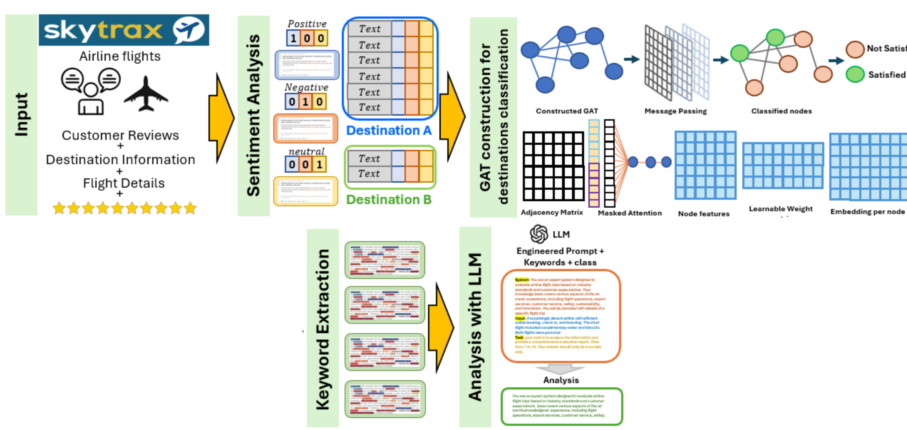

Fig. 1: The illustrative architecture of AIRNODE framework.

5 Experiment and Results

This section outlines the construction and execution of our experiment using the AIRNODE framework to analyze airline review data for enhancing passenger satisfaction. We describe data collection and preprocessing, feature engineering, graph construction, and the training and evaluation of models, focusing on their performance in classifying overall flyer satisfaction. Additionally, we present case studies that demonstrate the practical application of AIRNODE in identifying key aspects influencing passenger satisfaction and areas needing improvement.

5.1 Data Collection and Preprocessing

The dataset consists of over 128,000 airline reviews extracted from SkyTrax's Air Travel Review Website by Ljungström [15]. Each entry includes review text, ratings for various aspects of the flight experience, and metadata such as departure and destination codes. The primary label for prediction is the Recommended column, indicating whether a passenger recommended the journey.

Preprocessing steps included calculating journey distances based on geographic coordinates or lookup tables, standardizing dates to YYYY-MM-DD format, and handling missing values through imputation or removal if excessive.

5.2 Feature Engineering

Textual embeddings were generated using a pre-trained RoBERTa model, converting review text into dense vector representations. Other techniques such as BERT, TFIDF, and Word2Vec were also employed. Sentiment analysis was performed using NLTK's Sentiment Intensity Analyzer, with sentiment scores incorporated as features. Mean ratings for various attributes were computed and normalized using min-max normalization.

5.3 Graph Construction

Nodes in the graph represent unique destinations, while edges represent journeys between locations, with edge weights corresponding to normalized journey distances. The adjacency matrix was constructed to represent the graph structure, ensuring symmetry.

5.4 Model Training and Evaluation

Three graph-based models were evaluated: Graph Convolutional Network (GCN), Graph Convolutional Long Short-Term Memory (GCLSTM), and Graph Attention Network (GAT). Each model was trained using the Adam optimizer. Performance was evaluated using accuracy, precision, recall, and F1-score, with metrics computed for both training and test sets for each fold and final performance assessed by averaging these metrics across all folds.

The primary label for classification is overall flyer satisfaction with the destination, derived from aggregated reviews. Destinations with a model output score of 0.5 or higher are classified as "satisfied" (1), while those below 0.5 are classified as "not satisfied" (0).

This threshold helps distinguish destinations with higher satisfaction scores from those with lower scores, providing valuable insights for airlines to enhance their services and improve customer satisfaction.

The summarized binary classification algorithm that classifies the aggregated customer satisfaction for destination nodes is presented in Algorithm 1.

Algorithm 1 Graph Construction and Analysis

Input: Dataset D with reviews, departures, and destinations **Output:** Binary classification of flyer satisfaction with destinations **Initialize:** Graph $G = (V, E)$

procedure DATA PREPROCESSING(D)
  Calculate journey distance, handle missing values
end procedure
procedure GRAPH CONSTRUCTION(D)
  for each destination v_i in D **do**
    Add v_i to V with features $\mathbf{x}i$
  end for
  for each journey (v_i, v_j) in D **do**
    Add edge (v_i, v_j) with weight wij to E
  end for
  Construct adjacency matrix A from G
end procedure
procedure DEFINE GAT MODEL
  Define GAT model layers and constraints
end procedure
procedure MODEL TRAINING AND EVALUATION(D)
  Initialize StratifiedKFold with 10 folds
  for each fold f in StratifiedKFold **do**
    Train and evaluate GAT model on D_{train}
    Compute performance metrics on D_{test}
    Classify nodes as satisfied (1) or not satisfied (0)
  end for
  Compute average metrics across all folds
end procedure
Return Binary classification of flyer satisfaction

5.5 Results

In this section, we present the performance of various graph-based models—GCN, GAT, and GCLSTM—applied to airline review data. GCN was used to capture local node features and their immediate neighbors, providing a baseline for graph-based learning. GAT dynamically assigned different importance to nodes in a neighborhood, enhancing the model's capacity to learn complex relationships. GCLSTM combined the strengths of GCN with LSTM, capturing both spatial and temporal dependencies [18]. Each model was evaluated using different types of word embeddings combined with sentiment analysis and additional ratings (Wifi, Entertainment, Food, Ground Service, Comfort, and Service). Evaluation metrics included Accuracy, Precision, Recall, and F1 score, with a primary focus on test performance to determine the best model configuration. Additionally, two case studies were conducted where keywords were extracted from highly visited and less visited nodes, analyzing the results using a large language model.

Table 1 presents the performance data for the GCN model with varying model configurations that handle different sets of features. The results indicate that the best performance for the GCN model is achieved with the configuration that includes Word2Vec embeddings, sentiment analysis, and additional ratings. This configuration yields the highest test accuracy of 0.7832, precision of 0.8101, recall of 0.7503, and F1 score of 0.7664. These results suggest that integrating additional ratings with word embeddings and sentiment analysis improves the model's ability to classify whether a destination is considered "satisfied."

Table 1: GCN Model Performance with Different Configurations. Configuration 1: Word2Vec Embedding Features | Configuration 2: Word2Vec Embedding and Sentiment Analysis Features | Configuration 3: Word2Vec Embedding, Sentiment Analysis, and Additional Ratings Features.

Configuration	Accuracy	Precision	Recall	F1 Score
1	0.7276	0.7427	0.6930	0.6833
2	0.7513	0.7257	0.7932	0.7550
3	0.7832	0.8101	0.7503	0.7664

Table 2 presents the performance of the GCN, GAT, and GCLSTM models using different types of word embeddings, combined with sentiment analysis and additional ratings.

Table 2: Performance comparison of various embeddings with GCN, GAT, and GCLSTM models. Embedding (Emb), Accuracy (A), Precision (P), Recall (R), and F1-score (F1)

Emb	**GCN**				**GAT**				**GCLSTM**			
	A	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**
Word2Vec	0.7832	0.8101	0.7503	0.7664	0.8098	0.8230	0.7969	0.8028	0.7520	0.7800	0.7200	0.7500
TFIDF	0.7673	0.7470	0.8043	0.7976	0.7608	0.7462	0.7883	0.7609	0.7410	0.7600	0.7300	0.7450
BERT	0.7154	0.8097	0.5526	0.6490	0.8128	0.8037	0.8175	0.8096	0.7890	0.8000	0.7700	0.7850
RoBERTa	0.7543	0.7810	0.7033	0.7338	0.8464	0.8730	0.8950	0.8820	0.7876	0.8092	0.7410	0.7736

The results demonstrate that the GCN model with Word2Vec embeddings achieves the highest accuracy of 0.7832, while the GAT model with RoBERTa embeddings outperforms all other configurations across all metrics, achieving the highest test accuracy of 0.8464, precision of 0.8730, recall of 0.8950, and F1 score of 0.8820. The GCLSTM model with RoBERTa embeddings also shows good performance, with a test accuracy of 0.7876 and an F1 score of 0.7736.

Among the evaluated models and configurations, the Graph Attention Network (GAT) model with RoBERTa embeddings demonstrates the best overall performance on the binary classification task of identifying whether an airline destination is classified as "satisfied" or "not satisfied." The dynamic weighting of neighboring nodes in the GAT model effectively captures the intricate relationships in the data, leading to superior classification outcomes and is thus selected as the classification model for the AIRNODE framework.

5.6 AIRNODE Framework Case Study

To understand the factors influencing passenger satisfaction at various airports, we analyzed airline reviews using keyword extraction and K-means clustering. Clusters with the highest satisfaction ratings were labeled as "positive," while those with the lowest ratings were identified as "negative." This section presents insights for Guangzhou Baiyun International Airport (CAN) and Portland International Jetport (PWM), focusing on China Southern Airlines and United

Airlines, respectively. Our approach leverages AIRNODE (Attention-based Insights for Reviewing Node Optimized Destinations and Experiences) to provide a comprehensive analysis of these reviews, modeling complex relationships and extracting actionable insights.

Guangzhou Baiyun International Airport (CAN) Guangzhou Baiyun International Airport (CAN) is highly rated based on airline reviews for China Southern Airlines. Table 3 summarizes the key positive and negative aspects identified using AIRNODE.

Table 3: Analysis for China Southern Airlines at CAN for Positive (P) and Negative (N) Clusters (C)

C	**Keywords & Occurrences**	**Recommendations**	**Reasoning**
P	◦ Ground staff: 11 ◦ Flight attendant: 12 ◦ Cabin crew: 10 ◦ Crew friendly: 6 ◦ Business class: 12 ◦ Inflight entertainment: 4 ◦ Comfortable seat: 3 ◦ Smooth check-in: 9 ◦ Guangzhou: 9	◦ Emphasize quality of staff ◦ Enhance business class services ◦ Maintain high in-flight service standards ◦ Ensure comfort and cleanliness ◦ Efficient check-in processes ◦ Promote key routes	◦ High service standards and friendly attitudes are essential ◦ Promoting business class features attracts high-paying customers ◦ Updating entertainment options enhances experience ◦ Regular maintenance ensures pleasant journeys ◦ Streamlined processes improve travel
N	◦ Delay: 3 ◦ Bad experience: 3 ◦ Staff: 3 ◦ Flight time: 2 ◦ Inflight entertainment: 2 ◦ Check-in difficulty: 1	◦ Improve punctuality ◦ Enhance customer service ◦ Streamline check-in process ◦ Upgrade inflight entertainment	◦ Frequent delays indicate a need for better scheduling ◦ Negative feedback about staff suggests comprehensive training ◦ Difficult check-in processes highlight the need for efficiency ◦ Dissatisfaction with inflight entertainment suggests upgrades

Positive interactions with staff, such as friendly ground staff and attentive flight attendants, and commendations for business class and in-flight entertainment underscore the importance of maintaining high service standards. Passengers particularly appreciated the smooth check-in process and comfortable seating, indicating that these areas significantly enhance the overall travel experience. Furthermore, the emphasis on professionalism and the positive reputation of key routes like Guangzhou to Los Angeles contribute significantly to high passenger satisfaction.

However, frequent mentions of delays, bad experiences, and issues with staff highlight critical areas needing improvement. Negative feedback about flight times and difficulties during check-in processes indicate operational inefficiencies that need addressing. Additionally, the dissatisfaction with inflight entertainment suggests that upgrades and diversification of options are necessary.

While positive feedback highlights the excellence of staff interactions and business-class services, negative feedback about customer service suggests inconsistency. However, considering that there are more mentions of staff being good than bad in Table 3, it can be considered that generally, the staff operates professionally and satisfactorily with the case of a few exceptions.

Portland International Jetport (PWM) Portland International Jetport (PWM) is identified as a less recommended node based on airline reviews for

United Airlines. Table 4 summarizes the key positive and negative aspects identified using AIRNODE.

Table 4: Analysis for United Airlines at PWM for Positive (P) and Negative (N) Clusters (C)

C	Keywords & Occurrences	Recommendations	Reasoning
P	◦ Lovely experience: 1 ◦ Thoughtful service: 1 ◦ Upbeat staff: 1 ◦ Flight attendant: 1 ◦ Patience with boarding: 1 ◦ Infant care: 1	◦ Focus on thoughtful and upbeat customer experiences ◦ Maintain quality of flight attendants' service ◦ Ensure well-managed services for families	◦ Passengers value personal touches and positive interactions, enhancing loyalty ◦ Positive mentions of flight attendants suggest their service is significant ◦ Special care for families attracts more families
N	◦ Customer service: 7 ◦ Flight delays: 6 ◦ Maintenance issues: 2 ◦ Rude staff: 4	◦ Enhance customer service ◦ Reduce flight delays ◦ Improve communication ◦ Increase maintenance efficiency	◦ Consistent complaints about poor customer service highlight a critical area for improvement ◦ Frequent mentions of delays suggest inefficiencies ◦ Poor communication during delays is common ◦ Technical issues leading to delays indicate a need for better maintenance

Positive interactions with flight attendants and commendations for family-friendly services underscore the importance of maintaining high service standards. Passengers noted the lovely and thoughtful experiences provided by the airline staff, which significantly enhanced their travel satisfaction. Special care for families, particularly those with infants, was highly appreciated. The patience shown by staff during boarding processes also contributed positively to the overall experience.

However, PWM displays clashing trends. On one hand, there are commendations for individual staff interactions and family-friendly services, suggesting a potential strength in personal service. On the other hand, frequent mentions of poor customer service and delays highlight critical areas needing improvement. Complaints about rude staff and maintenance issues suggest significant gaps in training and operational efficiency. The higher frequency of mentions of rude staff indicates that this area needs improvement and the one mention of a good positive experience in Table 4 is an exception. Poor communication during delays was a common issue, indicating a need for better handling of such situations. The frequent mention of flight delays and technical issues suggests that operational inefficiencies are a significant problem that needs addressing.

Despite positive feedback on individual staff interactions and family-friendly services, the inconsistency in customer service quality is a major concern. The high frequency of complaints about delays and maintenance issues points to significant operational inefficiencies. Addressing these inconsistencies and improving overall operational efficiency is vital for enhancing passenger satisfaction.

6 Conclusion

This paper introduced AIRNODE (Attention-based Insights for Reviewing Node Optimized Destinations and Experiences), a two-stage model leveraging Graph

Attention Networks (GAT) to analyze airline review data for enhancing passenger satisfaction. By integrating RoBERTa embeddings with sentiment analysis and additional ratings, AIRNODE effectively captures complex relationships within the review data. Empirical evaluation demonstrated that the GAT model outperformed other graph-based models such as GCN and GCLSTM, achieving the highest accuracy, precision, recall, and F1 scores. The AIRNODE framework enables detailed analysis of destinations to identify aspects that contribute to passenger satisfaction and those that require improvement. Case studies on Guangzhou Baiyun International Airport (CAN) and Portland International Jetport (PWM) highlight key strengths and weaknesses, guiding airlines on maintaining and improving service quality.

AIRNODE provides a robust framework for leveraging large-scale review data to drive improvements in the airline industry, advancing the practical applications of graph-based machine learning in enhancing customer satisfaction. However, the study is limited by the scope and quality of the available data. Biases in the dataset, such as the tendency for more vocal customers to leave reviews, can skew the analysis. Additionally, lower quantities of data for specific nodes and journeys can lead to inconclusive results, reducing the reliability of insights for less frequently reviewed destinations. These limitations underscore the need for more comprehensive and balanced data collection to enhance the accuracy and generalizability of the findings.

References

1. Alanazi, M.S.M., Li, J., Jenkins, K.W.: Multiclass sentiment prediction of airport service online reviews using aspect-based sentimental analysis and machine learning. Mathematics **12**(5), 781 (2024)
2. Alegre, J., Garau, J.: Tourist satisfaction and dissatisfaction. Annals of Tourism Research **37**(1), 52–73 (2010)
3. Aljedaani, W., Rustam, F., Mkaouer, M.W., Ghallab, A., Rupapara, V., Washington, P.B., Lee, E., Ashraf, I.: Sentiment analysis on twitter data integrating textblob and deep learning models: The case of us airline industry. Knowledge-Based Systems **255**, 109780 (2022)
4. Çallı, L., Çallı, F.: Understanding airline passengers during covid-19 outbreak to improve service quality: topic modeling approach to complaints with latent dirichlet allocation algorithm. Transportation research record **2677**(4), 656–673 (2023)
5. Cambria, E., White, B.: Jumping nlp curves: A review of natural language processing research. IEEE Computational Intelligence Magazine **9**(2), 48–57 (2014)
6. Chen, C.F., Tsai, D.C.: How destination image and evaluative factors affect behavioral intentions? Tourism Management **28**(4), 1115–1122 (2007)
7. Gitto, S., Mancuso, P.: Improving airport services using sentiment analysis of the websites. Tourism management perspectives **22**, 132–136 (2017)
8. Gupta, M., Kumar, R., Walia, H., Kaur, G.: Airlines based twitter sentiment analysis using deep learning. In: 2021 5th International Conference on Information Systems and Computer Networks (ISCON). pp. 1–6 (2021). https://doi.org/10.1109/ISCON52037.2021.9702502
9. Hawkins, B.: Claude 3 vs. gpt-4o: What's the difference? (2024), https://claude101.com/claude-3-vs-gpt-4o/, accessed: 2024-05-29

10. Homaid, M.S., Bisandu, D.B., Moulitsas, I., Jenkins, K.: Analysing the sentiment of air-traveller: A comparative analysis. International Journal of Computer Theory and Engineering **14**(2), 48–53 (2022)
11. Iddrisu, A.M., Mensah, S., Boafo, F., Yeluripati, G.R., Kudjo, P.: A sentiment analysis framework to classify instances of sarcastic sentiments within the aviation sector. International Journal of Information Management Data Insights **3**(2), 100180 (2023)
12. Jing, X., Chennakesavan, A., Chandra, C., Bendarkar, M.V., Kirby, M., Mavris, D.N.: Bert for aviation text classification. In: AIAA AVIATION 2023 Forum. p. 3438 (2023)
13. Kwon, H.J., Ban, H.J., Jun, J.K., Kim, H.S.: Topic modeling and sentiment analysis of online review for airlines. Information **12**(2), 78 (2021)
14. Liu, B.: Sentiment analysis and opinion mining, vol. 5. Morgan & Claypool Publishers (2012)
15. Ljungström, Joel: Mining the Skies: An Exploration of Airline Reviews using LDA (2023), Student Paper
16. Mandal, S., Maiti, A.: Rating prediction with review network feedback: a new direction in recommendation. IEEE Transactions on Computational Social Systems **9**(3), 740–750 (2021)
17. Nguyen, T.D.: An approach to improve the accuracy of rating prediction for recommender systems. Automatika **65**(1), 58–72 (2024)
18. Rozemberczki, B., Scherer, P., He, Y., Panagopoulos, G., Riedel, A., Astefanoaei, L., Kiss, O., Sarkar, R., Gupta, P., Shia, B., Pellionisz, A., Staud, J.: Pytorch geometric temporal: Spatiotemporal signal processing with neural machine learning models (2021), `https://arxiv.org/abs/2107.13444`
19. Song, M., Feng, Y., Jing, L.: A survey on recent advances in keyphrase extraction from pre-trained language models. In: Vlachos, A., Augenstein, I. (eds.) Findings of the Association for Computational Linguistics: EACL 2023. pp. 2153–2164. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023). `https://doi.org/10.18653/v1/2023.findings-eacl.161`, `https://aclanthology.org/2023.findings-eacl.161`
20. Srinivas, S., Ramachandiran, S.: Passenger intelligence as a competitive opportunity: Unsupervised text analytics for discovering airline-specific insights from online reviews. Annals of Operations Research **333**(2), 1045–1075 (2024)
21. Subroto, A., Christianis, M.: Rating prediction of peer-to-peer accommodation through attributes and topics from customer review. Journal of Big Data **8**(1), 9 (2021)
22. Sulu, D., Arasli, H., Saydam, M.B.: Air-travelers' perceptions of service quality during the covid-19 pandemic: Evidence from tripadvisor sites. Sustainability **14**(1), 435 (2021)
23. Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., Wang, G.: Text classification via large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 8990–9005 (2023)
24. Wang, L., Guo, W., Yao, X., Zhang, Y., Yang, J.: Multimodal event-aware network for sentiment analysis in tourism. IEEE MultiMedia **28**(2), 49–58 (2021). `https://doi.org/10.1109/MMUL.2021.3079195`
25. Wu, S., Gao, Y.: Machine learning approach to analyze the sentiment of airline passengers' tweets. Transportation Research Record **2678**(2), 48–56 (2024)
26. Zhang, Y., Wang, M., Ren, C., Li, Q., Tiwari, P., Wang, B., Qin, J.: Pushing the limit of llm capacity for text classification. arXiv preprint arXiv:2402.07470 (2024)