Summarization of Airline Customer Reviews using NLP Techniques

Chun Yen Lin, Hsin Yu Liao

ANLT 293 ON1 NLP for Data Science

Dr. Pramod Gupta

05/02/2025

Abstract

This project applies natural language processing techniques to analyze and summarize airline customer reviews. A sentence-level pipeline combines zero-shot topic classification, rule-based fallback logic, and sentiment analysis to extract key content and highlight emotional intensity. Using 17 predefined aspect categories, the system generates concise summaries and identifies the most emotionally charged sentence in each review. Results from sample data demonstrate the system's ability to process both positive and negative feedback effectively. The approach provides a scalable method for transforming unstructured reviews into interpretable insights.

Keywords: NLP, Airline Reviews, Reviews Summarization, Zero-Shot Classification, Sentiment Analysis

Summarization of Airline Customer Reviews using NLP Techniques

Introduction

With the rapid growth of digital platforms, user-generated content (UGC) has become an important resource for both consumers and businesses. In the airline industry, reviews posted on platforms such as Skytrax, TripAdvisor, and airline-specific forums provide valuable insights into passengers' experiences across various service aspects, including punctuality, seating, staff behavior, and overall satisfaction.

Despite their usefulness, these reviews are typically written in free text, making it difficult to process them on a scale. A single review often covers multiple topics with varying emotional tones, and the sheer volume of available data presents challenges for both travelers seeking quick insights and airlines aiming to monitor service performance efficiently.

This project applies Natural Language Processing (NLP) techniques to analyze and summarize airline reviews at the sentence level. Using a combination of zero-shot classification with a transformer model (`facebook/bart-large-mnli`) (Lewis et al., 2020) and a rule-based fallback system, each sentence is assigned a relevant topic from a curated list of 17 airline service categories. Sentiment analysis is performed using a three-class model (`cardiffnlp/twitter-roberta-base-sentiment`) (Cardiff NLP, 2021) to detect emotional polarity, allowing the system to identify the most emotionally significant sentence in each review.

This project demonstrates the practical application of modern NLP tools in the context of customer feedback analysis. By combining topic classification and sentiment scoring, the system generates concise summaries that reflect both the thematic content and emotional highlights of customer reviews. This approach offers a scalable, interpretable solution for transforming unstructured feedback into structured insights, supporting better understanding of customer experiences in the airline sector.

Background

In the digital age, customer reviews have become a critical touchpoint for both travelers and service providers in the airline industry. Passengers regularly share detailed feedback online about their flight experiences, covering topics such as service quality, delays, staff behavior, and cabin conditions. These reviews offer valuable insights not only for prospective travelers making decisions, but also for airlines seeking to monitor performance and improve operations. However, the large volume of user-generated content, combined with its unstructured and subjective nature, makes manual analysis time-consuming and inefficient. This has led to growing interest in automated methods that can extract meaningful information from textual feedback in a scalable, interpretable way.

Methodology

Data Preparation

The dataset used in this project was sourced from Kaggle, titled “Airline Reviews.” It contains approximately 3,700 user reviews specifically related to British Airways (BA). Each entry includes a free-text customer review in the ReviewBody column.

To prepare the data for natural language processing, a series of cleaning and preprocessing steps were applied. First, all reviews containing missing values were removed to ensure data quality. Then, reviews with fewer than 10 characters were filtered out, as they were unlikely to contain meaningful content. Each remaining review was segmented into individual sentences using the *spaCy* library (Honnibal & Montani, 2021) with the `en_core_web_md` model and its built-in sentencizer. After segmentation, all text was converted to lowercase, and non-alphanumeric characters were stripped using regular expressions to ensure uniformity and eliminate noise. These steps produced a clean, sentence-level dataset suitable for classification and sentiment analysis tasks.

Topic Classification

To identify the main topics discussed in each sentence, this project adopted a hybrid classification approach combining zero-shot learning with rule-based fallback mechanisms. For the primary classification, the `facebook/bart-large-mnli` (Lewis et al., 2020) transformer model was used in a zero-shot setting. This model allowed each sentence to be evaluated against a predefined list of 17 airline-related topics—such as “Flight Delay,” “Cabin Service,” “In-flight Food and Drinks,” and “Value for Money”—without requiring any task-specific training. For each sentence, the model returned a ranked list of topic labels along with associated confidence scores.

However, as zero-shot models can be uncertain or unreliable when the confidence score is low, particularly in the presence of domain-specific phrases or implicit expressions, a secondary rule-based fallback system was implemented. This system consisted of manually crafted regular expressions that matched frequent keyword patterns associated with each topic. For instance, phrases like “screen” or “media system” would trigger the “In-flight Entertainment” label, even if the model’s confidence was low. The fallback system was applied only when the zero-shot model’s confidence score was below 0.3, ensuring that high-certainty predictions were preserved while low-confidence ones were intelligently overridden.

This layered classification strategy laid a strong foundation for the subsequent sentiment analysis, allowing both topic detection and emotional tone to be analyzed with greater precision.

Sentiment Analysis

To assess the emotional tone of each sentence, the `cardiffnlp/twitter-roberta-base-sentiment` model (Cardiff NLP, 2021) was used. This transformer-based model classifies text into one of three sentiment categories: Positive, Neutral, or Negative. This model outputs a probability distribution across the three sentiment classes. To streamline downstream processing and enable emotion-based comparisons between sentences, each predicted sentiment was mapped to a numerical polarity score: positive sentiment was assigned a score between 0 and +1, negative sentiment between 0 and -1, and neutral sentiment was treated as 0. This polarity score enabled the project to rank sentences by emotional intensity and extract the most emotionally charged sentence from each review. The emotional sentence was treated as a key representative of user satisfaction or dissatisfaction and was prioritized in the overall summarization strategy.

Summarization Selection Strategy

To generate meaningful summaries for each review, this project adopted an emotion-driven sentence selection strategy. After each sentence was classified and assigned a sentiment polarity score, the sentence with the highest emotional intensity—defined as the greatest absolute polarity score—was selected as the most emotional sentence. In cases where multiple sentences had similar intensity, preference was given to those discussing either “Overall Airline Experience” or “Value for Money,” as these categories most directly reflect the customer’s holistic satisfaction or perceived fairness.

Following the identification of the most emotional sentence, a summary was constructed by selecting up to three additional sentences that represented distinct high-confidence topics. These sentences were drawn from the previously classified results and were carefully filtered to exclude the most emotional sentence in order to reduce redundancy. Preference was given to topic sentences with classification confidence scores above a predefined threshold (0.25). If fewer than three high-confidence sentences were available, lower-confidence ones were used to ensure that the summary included multiple topics.

This combined approach ensured that each review summary captured both emotional highlights and thematic diversity, producing outputs that were concise, sentiment-aware, and representative of the review’s core content.

Output and Evaluation

The results of this project were displayed in the JupyterLab console using structured formatting. Each processed review included the original text, a system-generated topic summary of up to three distinct sentence-level insights, and the most emotionally salient sentence based on sentiment polarity. The summaries were designed to be human-readable and interpretable, clearly labeling each topic and associated sentence to enhance readability.

To facilitate documentation or further analysis, the system also supports optional export of the output data to a CSV file. The exported file includes the original review, topic summaries, corresponding topics, the most emotional sentence, and its sentiment score. This flexibility allows for both interactive, inline analysis during development and formal output archiving when needed.

Result

To evaluate the effectiveness of the NLP-based summarization system, the full processing pipeline was applied to a set of representative airline customer reviews sampled from the dataset. Each review was segmented into individual sentences, which were then classified into one of 17 predefined service-related topics using a *transformer-based zero-shot classifier* (Wolf et al., 2020). In cases where model confidence was low, a rule-based fallback mechanism was triggered to assign a more reliable topic based on domain-specific keywords. In parallel, sentiment polarity scores were computed using a three-class sentiment model to identify the most emotionally salient sentence within each review.

Review 1: Positive Experience in Club World on British Airways

This review reflects a highly satisfying Club World (Business Class) experience on a British Airways A350 flight. The reviewer praised almost every aspect of the flight, from seating and entertainment to food and crew professionalism.

Table 1

Sentence Classification Output for Review 1

Aspect Category	Representative Sentence	Confidence Score	Sentiment Score
Value for Money	An excellent flight in Club World on British Airways.	0.36	0.98
Staff & Service Attitude	The welcome aboard was warm and that continued throughout the flight.	0.23	0.96
Staff & Service Attitude	The crew were attentive, friendly and very professional.	0.50	0.95
In-flight Food and Drinks	On board food for dinner and breakfast was good and there was a well chosen selection of wines.	0.57	0.98

In-flight Entertainment	In flight entertainment offered a great selection of films and audio.	0.60	0.97
Seat Comfort	The seat/flat bed was very comfortable - British Airways have done an excellent job in the design and comfort of the suites on board the A350.	0.32	0.98
Value for Money	I liked the sleek, minimalist design.	0.18	0.95
Overall Airline Experience	This flight showed that BA can be among the world’s best airlines.	0.36	0.97

Table 2

Model-Generated Review Summary for Review 1

Aspect Category	Representative Sentence
In-flight Entertainment	In-flight entertainment offered a great selection of films and audio.
In-flight Food and Drinks	On board food for dinner and breakfast was good and there was a well chosen selection of wines.
Staff & Service Attitude	The crew were attentive, friendly and very professional.
Most Emotionally Charged Sentence	This flight showed that BA can be among the world’s best airlines. Sentiment Score: 0.97

According to the table 1 and table 2, the system successfully extracted positive sentiments across several service dimensions, including entertainment, food and beverage quality, and crew professionalism. The sentence *“This flight showed that BA can be among the world’s best airlines”* was identified as the most emotionally charged, with a sentiment score of 0.97, reflecting strong customer satisfaction. The final summary included key aspects such as “In-flight Entertainment,” "In-flight Food and Drinks," and "Staff & Service Attitude," demonstrating the system's ability to generate concise, multi-aspect summaries that reflect a positive overall experience.

Review 2: Aging Aircraft and Poor Value on Long-Haul Flight

This review, focused on a long-haul route from Buenos Aires to London Heathrow, expressed dissatisfaction with outdated equipment, inconsistent food service, and overall value for money.

Table 3

Sentence Classification Output for Review 2

Aspect Category	Representative Sentence	Confidence Score	Sentiment Score
Overall Airline Experience	Buenos Aires to London Heathrow return.	0.21	0.00
Overall Airline Experience	The aircraft is very old, cabin configuration is very old and tired.	0.22	-0.93
In-flight Entertainment	IFE screens have not been changed since they were first installed.	0.61	0.00
In-flight Entertainment	My iPod has a larger and more responsive screen. *	0.55	0.90
In-flight Food and Drinks	Before taking off in Buenos Aires, some pax, but not all, were offered water or orange juice.	0.44	0.00
Value for Money	I never got any.	0.13	-0.67
In-flight Food and Drinks	After take off, drinks were offered, followed by a hot meal.	0.58	0.00
In-flight Food and Drinks	Food choices ran out in the first row. *	0.55	-0.63

Seat Comfort	Seats were uncomfortable, footrests were jammed. *	0.63	-0.95
In-flight Food and Drinks	On the return flight, sparkling wine and water were offered before take off, followed by drinks and the meal I had chosen online was a beef stew with mashed potatoes.	0.53	0.00
In-flight Entertainment	Poor movie choices, miniature screen and uncomfortable seats.	0.42	-0.94
Staff & Service Attitude	Crew OK.	0.21	0.00
Flight Delay	No indication as to which toilets to use, either forward in business class or rear economy.	0.13	0.00
Value for Money	Having flown Norwegian on their B787 in their premium cabin on the same route, BA is a waste of my money. *	0.56	-0.82

* Indicates sentences where topic labels were assigned using the fallback rule system rather than the zero-shot classifier.

Table 4

Model-Generated Review Summary for Review 2

Aspect Category	Representative Sentence
Seat Comfort	Seats were uncomfortable, footrests were jammed.
In-flight Entertainment	IFE screens have not been changed since they were first installed.
In-flight Food and Drinks	After take off, drinks were offered, followed by a hot meal.

Most Emotionally Charged Sentence	The aircraft is very old, cabin configuration is very old and tired. Sentiment Score: -0.93

Based on the table 4, this system correctly identified six distinct topics, including “Seat Comfort”, “In-flight Entertainment”, and “Value for Money”. Several fallback rules were triggered to ensure proper topic assignment, particularly for complaints expressed through indirect language. The most emotionally negative sentence, *“The aircraft is very old, cabin configuration is very old and tired,”* received a polarity score of -0.93 and was appropriately classified under “Overall Airline Experience.”

Review 3: Poor Staff Behavior and Overpricing

This review highlighted severe dissatisfaction with the cabin crew and perceived overcharging.

Table 5

Sentence Classification Output for Review 3

Aspect Category	Representative Sentence	Confidence Score	Sentiment Score
Staff & Service Attitude	The staff are very rude and not trained properly.	0.65	-0.98
Staff & Service Attitude	No exceptions are made for children and elderly people.	0.10	0.00
Value for Money	The price of the ticket is very expensive given the distance and the service is extremely poor. *	0.57	-0.98

* Indicates sentences where topic labels were assigned using the fallback rule system rather than the zero-shot classifier.

Table 6

Model-Generated Review Summary for Review 3

Aspect Category	Representative Sentence
Staff & Service Attitude	The staff are very rude and not trained properly.
Most Emotionally Charged Sentence	The price of the ticket is very expensive given the distance and the service is extremely poor. Sentiment Score: -0.98

Due to system constraints, the exclusion of emotionally charged sentences from the final summary and confidence score filtering, only one aspect-level sentence was included in the output for the third review. This outcome reflects the model's prioritization of high-confidence, non-redundant content.

These results illustrate the system's ability to analyze both positive and negative reviews by accurately identifying key service topics and emotionally significant sentences. The integration of topic classification, fallback rules, and sentiment scoring enabled the generation of concise summaries that reflect both content and tone. Overall, the approach has shown that it can effectively capture core customer feedback in a structured and interpretable format.

Future work

Looking forward, there are several promising avenues to further enhance the accuracy, usability, and real-world impact of this project. A primary area of improvement is to boost the precision and reliability of classification results. This can be achieved by addressing class imbalance problems through techniques like Synthetic Minority Over-sampling Technique (SMOTE), as well as by fine-tuning model performance using hyperparameter optimization methods such as grid search combined with cross-validation. Additionally, integrating a confidence score mechanism would provide users with better transparency, helping them interpret how certain the model is about its predictions.

To translate these improvements into a more accessible and practical tool, a valuable next step would be to build a browser-based application such as a Chrome extension, that embeds the trained NLP model. This tool could display real-time summaries and sentiment insights directly on airline review websites, saving users from having to read lengthy, repetitive reviews. Such a solution would not only enhance user experience but also extend the system's reach by offering instant, actionable insights in a convenient and intuitive format.

Conclusion

This project explored the use of natural language processing techniques to extract, classify, and summarize airline customer reviews in a structured and interpretable format. By leveraging transformer-based zero-shot classification and integrating a rule-based fallback system, the model was able to assign topic labels with greater confidence and relevance. Sentiment analysis using a three-class model further allowed the identification of emotionally salient sentences, enriching the summarization output with both thematic and emotional dimensions. The sentence-level processing approach ensured that even multi-topic reviews could be broken down and understood more precisely.

The results demonstrated the system's ability to handle a wide range of review tones, from highly positive to strongly negative, and to produce concise summaries that capture key service aspects such as staff behavior, in-flight amenities, and overall experience. In particular, the fallback mechanism played a crucial role in compensating for low-confidence predictions by using domain-specific keyword rules, while sentiment scoring helped highlight customer satisfaction or dissatisfaction with numerical clarity. Together, these components contributed to a hybrid architecture that balances generalization power with rule-based interpretability.

In practice, this work shows how unstructured customer feedback can be transformed into actionable insights. Travelers benefit from faster decision-making by accessing distilled summaries, while airlines can more efficiently detect recurring issues and service trends. Although the current implementation is limited to console output and selected examples, the underlying framework offers strong potential for real-world deployment, especially when integrated into user-facing tools. The project lays a solid foundation for future enhancements in usability, automation, and domain adaptability.

References

Anshul Chaudhary, & Muskan Risinghani. (2023). *Airline reviews*. Kaggle. https://doi.org/10.34740/KAGGLE/DS/4044107

Cardiff NLP. (2021). *Twitter-roBERTa-base for Sentiment Analysis*. Hugging Face. https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

Honnibal, M., & Montani, I. (2021). *spaCy 3: Industrial-strength Natural Language Processing in Python* [Software]. Explosion. https://doi.org/10.5281/zenodo.1212303

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7871–7880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). *Transformers: State-of-the-art natural language processing*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6