Empowering Airline Route Decisions with LLM-Generated Pseudo-Labels and Zero-Shot Review Prediction

Abdulaziz Alhamadani[1], Khadija Althubiti[2], Jianfeng He[3], Shailik Sarkar[3], Lulwah Alkulaib[4], Abdul Raheem Shaik[5], Seungwon (Shawn) Lee[6], Mahmood Khan[2], and Chang-Tien Lu[3]

[1] Department of Data Science and Business Analytics, Florida Polytechnic University, Lakeland FL 33805, USA aalhamadani@floridapoly.edu
[2] Department of Hospitality and Tourism Management, Virginia Tech, Falls Church, VA 22043 USA {kalthubiti,Mahmood}@vt.edu
[3] Department of Computer Science, Virginia Tech, Falls Church, VA 22043 USA {jianfenghe, shailik, ctlu}@vt.edu
[4] Department of Computer Science, Kuwait University, Kuwait lalkulaib@cs.ku.edu.kw
[5] Bradley Department of Electrical and Computer Engineering, Virginia Tech, Falls Church, VA 22043 USA ashaik97@vt.edu
[6] Tourism and Events Management, George Mason University, Manassas, VA, USA slz@gmu.edu

Abstract. The airline industry has suffered a severe impact due to the COVID-19 pandemic. It resulted in significant financial losses. Strategic route planning is now an urgent need to mitigate the ongoing crisis. Motivated by the importance of customer sentiment in informing airline route decisions, this paper presents EAGLE (Enhancing Airline Groundtruth Labels and rEview rating prediction), a novel two-stage framework that leverages the power of Large Language Models (LLMs) to address the limitations of current works, which often rely on manual labeling and traditional machine learning models. In the first phase, EAGLE introduces a pseudo-labeling approach using LLMs to automatically label customer reviews to reduce the need for manual annotation and mitigate potential biases that exist in human labeling. The second phase employs a zero-shot LLM-based text classification method to predict customer sentiment and preferences from online reviews to provide a more accurate and context-aware analysis of customer feedback. Through extensive experiments, we demonstrate the effectiveness and robustness of EAGLE to demonstrate its superior performance compared to existing techniques. The proposed framework empowers airline companies to make data-driven decisions about route expansions, considering customer preferences and sentiments. Our contribution fibs in enhancing the objectivity of sentiment analysis and providing a comprehensive and scalable solution for airline route planning in the post-pandemic era, eventually leading to improved customer satisfaction and optimized operations.

Keywords: Airline industry · Large Language Model · Text Generation · Social media data mining

1 Introduction

The COVID-19 pandemic has had a profound impact on the aviation industry, causing unprecedented disruptions and challenges. As countries enforced border closures and restrictive measures to contain the spread of the virus, the demand for air travel plummeted, leading to a significant drop in the number of flights [4, 22]. According to the International Air Transport Association (IATA), global passenger traffic declined by 65.9% in 2020 compared to 2019, with international passenger demand falling by 75.6% [2]. Airlines suffered net losses of 126.4 billion USD on a revenue loss of 373 billion USD in 2020. Direct aviation jobs decreased by approximately 43%, and aviation-supported jobs are estimated to have reduced by 52% [2, 27].

The pandemic's impact on the aviation industry has been fast and severe, with many businesses struggling to survive [3, 6, 24]. Airlines have been forced to ground their fleets, lay off employees, and seek government support to stay afloat [4]. For example, Virgin Australia, one of Australia's largest airlines, entered voluntary administration in April 2020 due to the financial strain caused by the pandemic [3]. Similarly, numerous airlines and aviation-related businesses have faced bankruptcy or have had to significantly downsize their operations to cope with the crisis.

Amidst the turbulence caused by the pandemic, airlines have been compelled to reevaluate their strategic frameworks considering unprecedented challenges. The aviation industry has been actively adopting diverse strategies to navigate uncertain and increasingly complex business environments. Studies such as those by Linden [14] and Schwenker and Wulf [21] have employed simulations and scenario-based strategic planning to project multiyear trajectories, highlighting the significance of scenario analysis and dynamic decision-making in response to market shifts. Traditional strategic planning methods, including customer data collected from surveys, continue to play a pivotal role. The accessibility and rapid analysis of big data, particularly user-generated content (UGC), are also recognized as reliable and valid methods for informing strategic decision-making, complementing other internal and external determinants.

In the competitive airline industry, understanding customer preferences is critical for optimizing routes, services, and overall satisfaction [10,31,34]. Airlines traditionally relied on surveys to gather feedback, primarily for route expansion within existing markets (e.g., gauging interest in adding flights to nearby cities) [8]. However, this approach has limitations. Surveys often have low response rates and may not capture the full spectrum of customer sentiment. Other works relied heavily on traditional machine learning methods or sentiment analysis to evaluate the quality of service by evaluating the customer reviews on social media [1,5,5,7,9,11–13,19,29,30,32] . While these methods provide some insights, they often struggle to capture the nuanced and contextual information present in customer feedback, limiting their effectiveness in informing strategic decisions.

In our work, we propose using Large Language Models (LLMs) to bridge this gap by leveraging their ability to understand and generate human-like text, enabling more accurate and context-aware analysis of customer sentiment.

In this paper, we present EAGLE (Enhancing Airline Groundtruth Labels and rEview rating prediction), a novel two-stage framework that leverages the power of LLMs to address the limitations of traditional approaches in analyzing customer reviews for the airline industry. EAGLE introduces a pseudo-labeling approach using LLMs to automatically label customer reviews, reducing the need for manual annotation and mitigating potential biases that exist in human labeling. By employing zero-shot learning, our method enables the analysis and rating of reviews without requiring pre-defined examples for each rating category. In the second stage, EAGLE utilizes a zero-shot LLM-based text classification model to predict customer sentiment and preferences from online reviews. This approach allows for a more accurate and context-aware analysis of customer feedback, capturing the nuanced and contextual information that traditional methods often struggle to identify. We conduct extensive experiments to validate the effectiveness and robustness of EAGLE, demonstrating its superior performance compared to existing techniques. Our work has significant implications for the airline industry, enabling airlines to make more informed decisions, improve customer satisfaction, and optimize their route planning strategies in the post-pandemic era. By providing a comprehensive and scalable solution for customer review analysis, EAGLE paves the way for enhanced decision-making and strategic planning in the highly competitive airline industry. Our contributions are summarized as follows:

- We introduce a pseudo-labeling approach using Large Language Models (LLMs) to automatically label customer reviews, reducing the need for manual annotation and mitigating potential biases that exist in human labeling.
- We propose a zero-shot LLM-based text classification model for predicting customer sentiment and preferences from online reviews, enabling more accurate and context-aware analysis of customer feedback.
- We conduct extensive experiments to validate the effectiveness and robustness of EAGLE, demonstrating its superior performance compared to existing state-of-the-art techniques.

2 Related Work

2.1 Pseudo-Labeling and Semi-Supervised Learning

Annotating data for domain-specific natural language tasks is a challenge that has been addressed through pseudo-labeling and semi-supervised learning. Moezzi et al. [20] approached pseudo-labeling by utilizing an uncertainty-aware framework and observed that the selection of prediction with low uncertainty improves generalization. In a recent work [17], a large language model was used for a comparative study on the effectiveness of LLM-generated pseudo labels compared to human annotation for domain-specific text classification tasks. Although their result indicated superior performance on human-annotated data, the goal of the

project was to generate synthetic data for samples in under-represented classes which does not address the efficacy of generating labels for existing texts based on human-annotated examples or the inherent bias that exists in a small sample size of annotated data. Wang et al. [28] adapt a zero-shot learning approach that utilizes label phrase expansion to find more semantically aligned words or phrases for alleviating the class imbalance problem of the naive-zero shot approach. However, the quality of pseudo-labeled data is not generalizable to all human annotated dataset. Another work [33], addresses this issue by proposing a prototype-guided module that chooses samples around a prototype text representation of an under-labeled class to assign pseudo labels. However, this work does not leverage LLMs for assigning pseudo-labels to examine if the same issue persists. Despite useful insights gained from the mentioned approaches, there still exists a gap in harnessing the power of LLMs for pseudo-labeling. Hence, in this work, we leverage the power of pre-trained LLMs for zero-shot learning of pseudo-labels in airline customer reviews.

2.2 Text Classification Techniques for Customer Reviews

The field of text classification for customer reviews has advanced significantly, with recent works addressing various challenges. Sun et al. [26] introduced Clue And Reasoning Prompting (CARP), using progressive reasoning to enhance classification performance in large-scale language models. Similarly, Zhang et al. [35] presented RGPT, an adaptive boosting framework leveraging ensembling techniques for specialized text classification, outperforming state-of-the-art models. Nguyen et al. [18] integrated sentiment analysis and rating prediction to improve recommendation systems, while Subroto et al. [25] focused on predicting review ratings through machine learning models analyzing attributes and topics. Mandal et al. [16] introduced review network feedback, incorporating customer interactions to enhance recommendation systems. Our work differs by utilizing LLMs for zero-shot pseudo-labeling and text classification, providing accurate and context-aware analysis of customer feedback without extensive manual annotation. This scalable and adaptable framework can be applied across various domains, including the airline industry, to drive customer-centric decision-making and improve business outcomes.

2.3 Customer Feedback Analysis in the Aviation Industry

Customer feedback analysis has gained significant attention in the aviation industry to improve services and inform decision-making. Various approaches, including sentiment analysis, topic modeling, and machine learning, have been employed. BERT, adapted for aviation-specific tasks, shows the potential of fine-tuned models for processing vast text data [12]. Sentiment analysis, widely applied to understand passenger opinions, compares methods like VADER and logistic regression [9]. Combining topic modeling and sentiment analysis helps identify key issues in airline reviews, highlighting factors like seat, service, meals, and delays [13].

Advanced techniques such as sarcasm detection [11], deep learning [7], multimodal approaches [30], time series methods [32], and aspect-based sentiment analysis [1] address the nuances of aviation customer feedback. These studies capture sudden changes in passenger sentiments, aiding airlines in taking mitigatory measures. Data from blogs and text extracting software assess service levels perceived by airport customers [5]. Despite these advancements, many rely on human-labeled data, which can introduce biases, and primarily focus on sentiment analysis.

Our method utilizes LLMs for zero-shot pseudo-labeling, reducing the need for human labeling and associated biases. We conduct multi-class text classification on user ratings, offering a granular understanding of customer feedback. These diverse applications and our novel approach enhance decision-making and service quality in the aviation industry.

3 Methodology

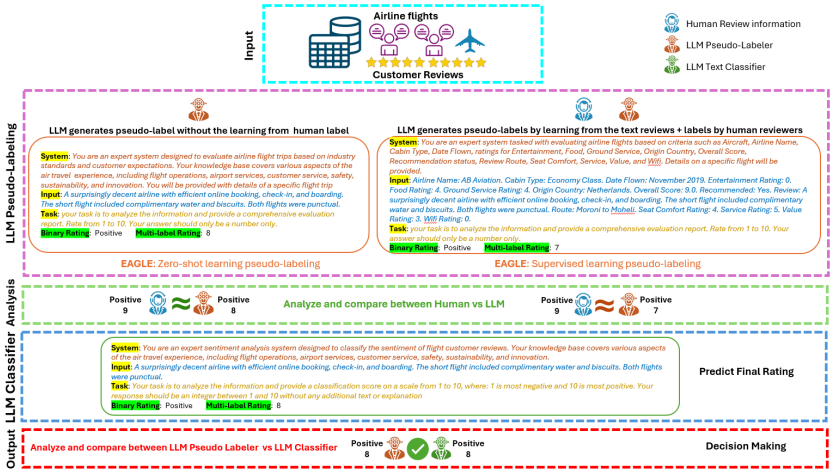

Fig. 1: The illustrative architecture of EAGLE framework.

In this study, we explain our approach (illustrated in figure. 1) which splits into two main parts: first, we develop an LLM pseudo-labeling to give an initial rating to customer reviews, and second, we classify these initially rated reviews to predict their final ratings. This two-step process helps us better understand and predict how customers feel about airline services for better decision-making. Our method uses two types of learning: zero-shot and supervised learning. Zero-shot learning helps us analyze and rate reviews without needing examples of each rating beforehand, while supervised learning uses examples with known ratings to learn how to rate new reviews accurately. We create specific instructions, or prompts, for the LLM to follow so it can understand and rate the text reviews on a scale of 1 (*most negative*) to 10 (*most positive*).

3.1 Pseudo-labeling with LLM

Notation: Let $A_G = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N_G}$ be the dataset contains the text reviews labeled by human with N_G

Define $A_G = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N_G}$ as the dataset comprising human-labeled text reviews, where N_G denotes the total number of samples. Each input $x^{(i)}$ represents a text review, and its corresponding label $y^{(i)} = [y_1^{(i)}, \ldots, y_C^{(i)}] \in \{1, \ldots, 10\}^C$ encapsulates the ratings across C class categories. In this rating system, $y_c^{(i)} = 1$ signifies the most negative review sentiment, and $y_c^{(i)} = 10$ signifies the most positive review sentiment. The objective is to generate a pseudo-label $\tilde{y}^{(i)}$ for every $y^{(i)}$ by learning from the LLM on the dataset A_G. The output will be the updated dataset $A_S = \{(x^{(i)}, y^{(i)}, \tilde{y}^{(i)})\}_{i=1}^{N_S}$

In this scenario where the pseudo-labeling is performed using a LLM, the adaptation to the specific pseudo-labeling task can be achieved through methods such as:

- Zero-shot Learning: The LLM is directly queried with inputs phrased in a way that includes the task description, aiming for the model to leverage its pre-trained knowledge to generate pseudo labels.
- Supervised Learning: For pseudo labeling, the LLM is first trained on a labeled dataset specific to the task. It then uses this trained model to generate pseudo labels for the dataset.

Zero-shot Learning Pseudo-labeling Prompt Construction The prompt, shown as $x_{\text{prompt}}^{(i)}$, is made using the variable $x^{(i)}$ and includes three main parts:

(1) System Description: x_{system} succinctly encapsulates the operational essence of what the LLM can do. It describes the wide range of information the model knows about and how it uses this knowledge to look at different kinds of data. For pseudo-labeling, we use this example prompt: *You are an expert system designed to evaluate airline flight trips based on industry standards and customer expectations. Your knowledge base covers various aspects of the air travel experience, including flight operations, airport services, customer service, safety, sustainability, and innovation. You will be provided with details of a specific flight trip.*

(2) Input: x_{input} represents the textual sequence intended for pseudo-labeling. The composite prompt, x_{prompt}, is methodically constructed by the concatenation of the system description x_{system}, the input x_{input}, and the task directive x_{task}. Formally, this can be represented as:

$$x_{\text{prompt}} = x_{\text{system}} \mid\mid x_{\text{input}} \mid\mid x_{\text{task}}$$

where $||$ denotes the concatenation operation.

(3) Task Description: x_{task} explains the objective to be accomplished, aligning the task of pseudo-labeling with the context of the LLM's capabilities. For pseudo-labeling, the task is explained as: *your task is to analyze the information and provide a comprehensive evaluation report. Rate from 1 to 10. Your answer should only be a number only.*

Supervised Learning Pseudo-labeling Prompt Construction The prompt, denoted as $x^{(i)}_{prompt}$, is built using the variable $x^{(i)}$ and comprises three distinct elements:

(1) System Description: x_{system} concisely outlines the capabilities of the LLM, detailing its extensive knowledge base and application across various data types. For pseudo-labeling, consider the following example: *You are an expert system tasked with evaluating airline flights based on criteria such as Aircraft, Airline Name, Cabin Type, Date Flown, ratings for Entertainment, Food, Ground Service, Origin Country, Overall Score, Recommendation status, Review Route, Seat Comfort, Service, Value, and Wifi. Details on a specific flight will be provided.*

(2) Input: x_{input} is the text sequence for pseudo-labeling, highlighting features and customer ratings in the reviews. Example input text: *Airline Name: AB Aviation. Cabin Type: Economy Class. Date Flown: November 2019. Entertainment Rating: 0. Food Rating: 4. Ground Service Rating: 4. Origin Country: Netherlands. Overall Score: 9.0. Recommended: Yes. Review: A surprisingly decent airline with efficient online booking, check-in, and boarding. The short flight included complimentary water and biscuits. Both flights were punctual. Route: Moroni to Moheli. Seat Comfort Rating: 4. Service Rating: 5. Value Rating: 3. Wifi Rating: 0.* The full prompt, x_{prompt}, is created by concatenating x_{system}, x_{input}, and x_{task} as follows:

$$x_{\text{prompt}} = x_{\text{system}} \mid\mid x_{\text{input}} \mid\mid x_{\text{task}}$$

where $||$ symbolizes the joining of these texts.

(3) Task Description: x_{task} outlines the goal, tailoring the pseudo-labeling activity to the LLM's strengths. The task is detailed as: *Your job is to assign an overall rating between 1 and 10, based solely on a numerical response.*

3.2 Customer Reviews Text Classification with LLM

Notation: Let $A_S = \{(x^{(i)}, y^{(i)}, \tilde{y}^{(i)})\}_{i=1}^{N_S}$ be the dataset containing the text reviews labeled by LLM (pseudo labels) with N_S samples, where $x^{(i)}$ is a text review input, and $\tilde{y}^{(i)} = [\tilde{y}_1^{(i)}, \ldots, \tilde{y}_C^{(i)}] \subseteq \{1, \ldots, 10\}^C$ is the corresponding label with C class categories. For a sample i, $\tilde{y}_c^{(i)} = 1$ denotes that the text review rating is most negative, and $\tilde{y}_c^{(i)} = 10$ denotes that the text review rating is most positive. The objective is to accurately map the input text reviews $x^{(i)}$ to their corresponding labels $\tilde{y}^{(i)}$, leveraging the intrinsic capabilities of the LLM.

Given the utilization of an LLM for classification, the training or inference process diverges from conventional approaches that rely on minimizing a predefined loss function. Instead, the model leverages its pre-existing knowledge, acquired through extensive pre-training on diverse text corpora, to perform the task of text classification. The process can be formalized as follows:

$$\hat{y}^{(i)} =_c P(\tilde{y}_c^{(i)} \mid x^{(i)}; \theta)$$

where $x^{(i)}$ is a text review input, $\tilde{y}^{(i)}$ is the set of possible labels, $\hat{y}^{(i)}$ is the predicted label by the LLM, and θ represents the parameters of the LLM. The prediction $\hat{y}^{(i)}$ is determined by selecting the label c that maximizes the conditional probability $P(\tilde{y}_c^{(i)} \mid x^{(i)}; \theta)$, which is computed by the LLM.

In this scenario where the text classification is performed using a LLM, the adaptation to the specific classification task can be achieved through methods such as zero-shot Learning where the LLM is directly queried with inputs phrased in a way that includes the task description, expecting the model to leverage its pre-trained knowledge to infer the correct label without any task-specific fine-tuning.

Zero-shot Learning Text Classification Prompt Construction Given the dataset $A_S = \{(x^{(i)}, y^{(i)}, \tilde{y}^{(i)})\}_{i=1}^{N_S}$ for text classification using an LLM, the prompt construction for zero-shot Learning is outlined as follows, utilizing $x^{(i)}$ as the input:

(1) System Description: This part, $x_{\text{system}}^{(i)}$, provides a brief on the LLM's design and domain expertise. For instance, *You are an expert sentiment analysis system designed to classify the sentiment of flight customer reviews. Your knowledge base covers various aspects of the air travel experience, including flight operations, airport services, customer service, safety, sustainability, and innovation.*

(2) Input: Denoted as $x_{\text{input}}^{(i)}$, this component is the actual text review that needs classification. It is the raw data on which the LLM operates, without any additional processing or context added.

(3) Task Description: The task, $x_{\text{task}}^{(i)}$, clarifies the objective of the classification, structured as a directive to the LLM. An example task could be, *Your task is to analyze the information and provide a classification score on a scale from 1 to 10, where: 1 is most negative and 10 is most positive. Your response should be an integer between 1 and 10 without any additional text or explanation*

The prompt $x_{\text{prompt}}^{(i)}$ for each i^{th} instance in the dataset is formulated by amalgamating these elements, expressed as:

$$x_{\text{prompt}}^{(i)} = x_{\text{system}}^{(i)} \mid\mid x_{\text{input}}^{(i)} \mid\mid x_{\text{task}}^{(i)}$$

where $||$ signifies the concatenation operation. This structured prompt facilitates the LLM's zero-shot Learning capability by providing a clear, comprehensive context for each text classification task.

4 Experiment and Results

In this section, we present the dataset and the evaluation of our proposed framework. To assess the effectiveness of our approach, we conduct experiments in two distinct stages. First, we employ an LLM to generate pseudo-labels for the dataset under two different setups: binary pseudo-labeling and multi-class pseudo-labeling. Second, we evaluate the performance of the LLM in classifying customer reviews, considering both binary and multi-class classification tasks.

4.1 Dataset

The dataset used in this study was collected by Ljungström [15], who scraped and extracted public reviews from the Air Travel Review (ATR) website[7]. ATR is a customer forum owned and operated by the airline rewards company SkyTrax, providing comprehensive reviews for airports, airport lounges, airline seats, and airlines. The dataset contains individual reviews related to airlines on ATR, consisting of 21 variables. The full dataset comprises 128,631 reviews from 547 airlines, ranging from April 2005 to September 2022. However, due to the development of ATR over time, not all dimensions have been consistently available for rating, resulting in incomplete data for some variables. In our experiments, rows with missing values were excluded from the analysis. For certain parts of the experiment, we focus on two specific variables: *Review* (The customer feedback text on the taken trip) and *OverallScore* (the provided rating by the customer). In other parts, we include all 21 variables, such as AirlineName, DataFlown, CabinType, Recommended, etc. To address class imbalance, we randomly sample 1,700 reviews for each of the three categories based on OverallScore: "Low" (1-3), "Medium" (4-6), and "High" (7-10) which results in 5100 samples. The distribution of word counts for the Review variable ranges from 1 to 1,058 words, with an average length of 136 words.

4.2 Experiment Settings and Evaluation Metrics

In this research, we employed Claude, an AI model developed by Anthropic, for pseudo-labeling and text classification tasks. We specifically used Claude version "claude-3-opus-20240229" due to its superior performance compared to other language models like ChatGPT. Previous studies have demonstrated that Claude outperforms ChatGPT in terms of accuracy and consistency across various natural language processing tasks [23]. To optimize the performance of Claude for pseudo-labeling and text classification, we carefully tuned the hyperparameters of the Claude API. The engine name was set to "claude-3-opus-20240229", which represents the specific version of Claude used in our experiments. We configured the maximum number of tokens to 1000, allowing the model to generate sufficiently long responses while maintaining computational efficiency. The temperature parameter, which controls the randomness of the generated outputs, was set to 0 to ensure deterministic and consistent results. The default top_p parameter which is 1 was used which means all possible tokens were considered during the sampling process. To ensure the robustness and reliability of our results, we ran each prompt through the Claude API three times and calculated the average performance across the three runs. This approach helps mitigate potential variability in the model's outputs and provides a more stable and representative assessment of its performance.

[7] `https://www.airlinequality.com`

4.3 Pseudo-labeling Results and Analysis

In this experiment, we developed two zero-shot learning pseudo-labeling prompts, denoted as ZP1 and ZP2, and two supervised learning pseudo-labeling prompts, denoted as SP1 and SP2. ZP1 and SP2 are provided in the methodology section 3. ZP2 has less instructions and is a more simplified version of ZP1 in terms of the system and task. SP1 is also a simplified version of SP2 and less details and instructions in terms of system and tasks as well. The performance of these engineered prompts was evaluated on two levels: binary and multi-class pseudo-labeling.

For binary pseudo-labeling, we considered any label or rating $\tilde{y}$ above 5 as positive and any label or rating $\tilde{y}$ below 5 as negative. The effectiveness of the binary pseudo-labeling was measured using the Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) metrics. These metrics were calculated between the human-assigned labels or ratings and the pseudo-labels generated by the prompts.

For multi-class pseudo-labeling, the prompts provided labels or ratings on a scale from 1 to 10, with 1 being the most negative and 10 being the most positive. The performance of the multi-class pseudo-labeling was evaluated using the same metrics as in the binary case: MAE, MSE, and R^2. These metrics were calculated between the human-assigned labels or ratings and the multi-class pseudo-labels generated by the prompts.

	Pseudo-binary labeling			**Pseudo-multi-labeling**		
Model	**MAE**	**MSE**	R^2	**MAE**	**MSE**	R^2
ZP1	**0.1466**	**0.0344**	**0.7687**	**1.3417**	**2.7155**	**0.7746**
ZP2	0.1727	0.7575	0.2836	2.3249	11.4223	0.0249
SP1	0.3064	0.1807	-0.2176	2.8575	15.6064	-0.2961
SP2	0.2749	0.1604	-0.0788	2.7323	14.5931	-0.2109

Table 1: Comparison of the first stage of EAGLE's performance in pseudo-binary-labeling and pseudo-multi-label between zero-shot and supervised learning.

The performance of the four pseudo-labeling models, namely ZP1, ZP2, SP1, and SP2, was evaluated in both binary and multi-class settings, as presented in Table 1. In the binary pseudo-labeling task, ZP1 demonstrated the best performance with the lowest MAE of 0.1466 and MSE of 0.0344, as well as the highest R-squared (R^2) value of 0.7687. This indicates that ZP1 achieves the highest agreement with human annotations in the binary setting. ZP2 exhibited slightly lower performance compared to ZP1, with an MAE of 0.1727, MSE of 0.7575, and R^2 of 0.2836. On the other hand, the supervised learning models, SP1 and SP2, showed relatively higher error rates and lower R^2 values in the binary task, suggesting that they may not capture the binary labels as effectively as the zero-shot learning models.

In the multi-class pseudo-labeling task, ZP1 once again outperformed the other models, achieving the lowest MAE of 1.3417 and MSE of 2.7155, along with

the highest R^2 value of 0.7746. This suggests that ZP1 is able to generate multi-class pseudo-labels that align well with the human annotations. ZP2 exhibited higher error rates and a significantly lower R^2 value of 0.0249 in the multi-class setting, indicating a weaker agreement with human labels compared to ZP1. The supervised learning models, SP1 and SP2, showed even higher error rates and negative R^2 values in the multi-class task, implying that they may struggle to capture the nuances of the multi-class labels.

Overall, the results demonstrate that the zero-shot learning model ZP1 consistently outperforms the other models in both binary and multi-class pseudo-labeling tasks. It achieves the lowest error rates and highest R^2 values, indicating a strong agreement with human annotations. The supervised learning models, SP1 and SP2, exhibit relatively lower performance in both settings, suggesting that they may not be as effective in capturing the underlying label distribution. These findings highlight the potential of zero-shot learning approaches, particularly ZP1, for generating accurate pseudo-labels in both binary and multi-class scenarios.

k	**Pseudo-binary labeling**				**Pseudo-multi labeling**			
	MAE	**MSE**	**RMSE**	R^2	**MAE**	**MSE**	**RMSE**	R^2
1000	0.1567	0.0401	0.1992	0.7268	1.3715	2.9592	1.7202	0.7481
2000	0.1542	0.0381	0.1948	0.7467	1.3698	2.8560	1.6899	0.7632
3000	0.1513	0.0368	0.1918	0.7553	1.3633	2.8288	1.6819	0.7673
4000	0.1487	0.0357	0.1889	0.7608	**1.3413**	2.7379	1.6546	0.7724
5000	**0.1466**	**0.0344**	**0.1860**	**0.7687**	1.3417	**2.7155**	**1.6479**	**0.7746**

Table 2: Performance metrics for pseudo-binary labeling and pseudo-multi labeling tasks.

To optimize computational resources and efficiency, we investigated the minimum amount of data required to achieve a desired level of performance using the best-performing prompt, ZP1. Let $D = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ denote the entire dataset, where x_i represents the input features and y_i represents the corresponding human-annotated labels. We define a subset of the dataset as $D_k = (x_1, y_1), (x_2, y_2), ..., (x_k, y_k)$, where $k \in 1, 2, ..., n$ represents the size of the subset.

We conducted a series of experiments by gradually increasing the size of the subset D_k and evaluating the performance of ZP1 on each subset. The performance was measured using the MAE, MSE, and R^2 metrics, as defined in the previous section. Let $MAE(k)$, $MSE(k)$, and $R^2(k)$ denote the respective metrics calculated on the subset D_k.

By incrementally increasing the size of the subset D_k and evaluating the performance metrics, we aim to identify the minimum amount of data required to achieve the desired performance level. Table 2 presents the performance metrics for the best-performing prompt, ZP1, on subsets of the dataset with varying sample sizes. The metrics are evaluated for both pseudo-binary labeling and pseudo-multi-labeling tasks.

In the pseudo-binary labeling, as the number of samples increases from 1000 to 5000, we observe a consistent improvement in performance. The MAE decreases from 0.1567 to 0.1466, indicating a reduction in the average absolute difference between the predicted and true binary labels. Similarly, the MSE decreases from 0.0401 to 0.0344, suggesting a decrease in the average squared difference between the predicted and true labels. The RMSE also shows a decreasing trend, from 0.1992 to 0.1860, indicating a reduction in the standard deviation of the prediction errors. Moreover, the R^2 value increases from 0.7268 to 0.7687, indicating an improvement in the proportion of variance in the true labels that can be explained by the predicted labels.

In the pseudo-multi-labeling, we observe a similar trend of improving performance as the sample size increases. The MAE decreases from 1.3715 to 1.3417, indicating a reduction in the average absolute difference between the predicted and true multi-class labels. The MSE also decreases from 2.9592 to 2.7155, suggesting a decrease in the average squared difference between the predicted and true labels. The RMSE shows a decreasing trend, from 1.7202 to 1.6479, indicating a reduction in the standard deviation of the prediction errors. The R^2 value increases from 0.74805 to 0.7746, indicating an improvement in the proportion of variance in the true labels that can be explained by the predicted labels.

It is worth noting that the improvement in performance metrics becomes smaller as the sample size increases, suggesting a diminishing return in performance gain. For example, the decrease in MAE is more pronounced when the sample size increases from 1000 to 2000 compared to the decrease when the sample size increases from 4000 to 5000. This observation aligns with the concept of learning curves, where the performance improvement tends to plateau as the sample size becomes sufficiently large.

The results in Table 2 provide insights into the trade-off between sample size and performance. While increasing the sample size generally leads to better performance, the marginal improvement diminishes as the sample size grows larger.

4.4 Text classification Results and Analysis

In this experiment, we conducted zero-shot text classification and evaluated the results on two levels: binary and multi-label classification. For binary classification, any rating $\tilde{y}$ above 5 was considered *positive*, while any rating $\tilde{y}$ below 5 was considered *negative*. We measured various performance metrics, including MAE, MSE, RMSE, R^2, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the model's performance, considering different aspects such as accuracy, precision, and recall, which are essential for assessing the effectiveness of the classification model.

The dataset was divided into training, validation, and testing sets using cross-validation. We compared our work to traditional machine learning models (e.g., Logistic Regression, Naïve Bayes, SVM, Random Forest), deep learning models (e.g., CNN, RNN), and pre-trained models such as BERT and RoBERTa. RoBERTa was fine-tuned on an NVIDIA 3060 GPU with hyperparameters tuned

on the validation set, including learning rates $\in \{2e-5, 3e-5, 4e-5\}$, batch sizes $\in \{16, 32\}$, a dropout rate of 0.3, a weight decay of 0.01, and a warmup proportion of 0.01.

Similar to the pseudo-labeling experiment, we measured the minimum amount of data required to achieve a certain level of performance for both binary and multi-label classification. Table 3 presents the binary classification results, while Table 4 shows the multi-label classification results.

Table 3: Binary classification results

k	MAE	MSE	RMSE	R^2	P	R	F1
1000	0.0690	**0.0152**	**0.1235**	**0.8325**	0.99	0.97	0.97
2000	0.0695	0.0157	0.1255	0.8258	0.98	0.97	0.97
3000	0.0710	0.0173	0.1317	0.8058	0.98	0.96	0.97
4000	0.0686	0.0158	0.1259	0.8198	0.98	0.96	0.97
5000	**0.0686**	0.0154	0.1241	0.8237	0.99	0.97	0.97

For binary classification (Table 3), as the number of samples (k) increases, the performance metrics remain relatively stable. The MAE ranges from 0.0686 to 0.071, indicating a low average absolute error. The MSE and RMSE also show low values, suggesting a good fit of the model. The R-squared values are consistently high (above 0.8), indicating that the model explains a significant portion of the variance in the data. The precision, recall, and F1-score for the binary classification is high, demonstrating the model's ability to accurately classify both positive and negative samples.

k	MAE	MSE	RMSE	R^2	P	R	F1	Purity
1000	0.7960	3.5102	1.8735	0.3933	0.81	0.76	0.78	0.7643
2000	0.7622	3.2422	1.8003	0.4294	0.81	0.76	0.78	0.7644
3000	0.7624	3.2424	1.8007	**0.4296**	0.81	0.76	0.78	0.7675
4000	0.7911	3.4689	1.8625	0.3798	**0.82**	0.76	0.78	0.7675
5000	**0.7507**	**3.1767**	**1.7823**	0.4282	0.81	0.76	0.78	0.7644

Table 4: Multi-label classification results

In the multi-label classification (Table 4), as the number of samples (k) increases, the performance metrics show improvements. The MAE decreases from 0.796 to 0.7507, indicating a reduction in the average absolute error. The MSE and RMSE also decrease, suggesting a better fit of the model. The R-squared values increase, indicating that the model explains more variance in the data as the sample size grows. The precision, recall, and F1-score remain relatively stable, with values around 0.81, 0.76, and 0.78, respectively. We also corroborate our results by measuring the purity. Purity is a metric commonly used to evaluate the quality of clustering algorithms, including those used in multi-classification tasks. It measures the extent to which clusters contain a single class or category. The purity score is between 0.7643 and 0.7644 which indicates that, on average, 76.44% of the instances within each cluster belong to the same true class. This suggests that EAGLE achieves a reasonably good level of cluster homogeneity.

Model	Precision	Recall	F1-score	Accuracy
LR	0.50	0.57	0.47	0.57
NB	0.31	0.53	0.39	0.53
SVM	0.42	0.57	0.44	0.57
Random Forest	0.51	0.55	0.41	0.55
LR (BoW)	0.51	0.54	0.52	0.54
LR (TF-IDF)	0.52	0.57	0.47	0.57
LR (Word)	0.39	0.46	0.41	0.46
CNN	0.53	0.58	0.52	0.57
RNN	0.30	0.45	0.34	0.45
Bert	0.63	0.65	0.64	0.65
Roberta	0.63	0.64	0.64	0.65
EAGLE	**0.81**	**0.76**	**0.78**	**0.76**

Table 5: Model performance comparison.

Table 5 compares the performance of our model EAGLE with various baselines. EAGLE achieves the highest precision (0.81), recall (0.76), F1-score (0.78), and accuracy (0.76) among all the models, outperforming traditional machine learning models, deep learning models, and pre-trained models such as BERT and RoBERTa.

Class	Precision	Recall	F1-Score
1	0.4108	0.7984	0.5425
2	0.7206	0.7571	0.7384
3	0.9245	0.7696	0.8414
4	0.6829	0.7421	0.7112
5	0.1316	0.8065	0.2262
6	0.7130	0.7490	0.7306
7	0.7617	0.7920	0.7766
8	0.8825	0.7411	0.8056
9	0.2591	0.7576	0.3861

Table 6: Classification performance by class.

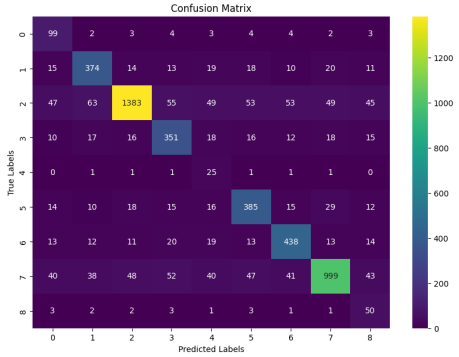

Fig. 2: Classification performance by class confusion matrix

Finally, Table 6 and the confusion matrix 2 show the performance of EAGLE for each class in the multi-label classification task. The precision, recall, and F1-score vary across different classes, with some classes (e.g., class 3 and class 8)

achieving higher performance compared to others (e.g., class 5 and class 9). This suggests that the model's performance may be influenced by the characteristics and distribution of the different classes in the dataset.

In summary, our zero-shot text classification model EAGLE demonstrates strong performance in both binary and multi-label classification tasks, outperforming various baseline models. The results highlight the effectiveness of our approach and its potential for accurate text classification with limited labeled data.

5 Conclusion

In this paper, we introduced EAGLE, a two-stage framework leveraging Large Language Models (LLMs) to enhance the analysis of airline customer reviews. EAGLE overcomes traditional limitations, offering deep insights into customer sentiment and preferences at scale. Our LLM-based pseudo-labeling automates labeling, reducing manual efforts and mitigating biases. Additionally, the zero-shot LLM-based text classification model accurately analyzes customer feedback, capturing subtle and contextual information often missed by traditional methods. Extensive experiments show EAGLE's superior performance over existing techniques, offering significant implications for the airline industry. EAGLE empowers airlines to make informed decisions, improve customer satisfaction, and optimize route planning, which is crucial for post-pandemic recovery and growth. Adopting EAGLE provides insights into customer preferences and feedback, guiding strategic decisions like expanding flight offerings or modifying routes. Future research can extend EAGLE to other domains and integrate additional data sources to further refine customer sentiment analysis.

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. COVID, I.I.U.: Financial impacts-relief measures needed-[press release] 2020 https://www. iata. org/en/pressroom/pr/2020-03-05-01. Retrieved from.[Google Scholar] (19)
3. Dube, K., Nhamo, G.: Major global aircraft manufacturers and emerging responses to the sdgs agenda. Scaling up SDGs Implementation: Emerging Cases from State, Development and Private Sectors pp. 99–113 (2020)
4. Dube, K., Nhamo, G., Chikodzi, D.: Covid-19 pandemic and prospects for recovery of the global aviation industry. Journal of Air Transport Management **92**, 102022 (2021)
5. Gitto, S., Mancuso, P.: Improving airport services using sentiment analysis of the websites. Tourism management perspectives **22**, 132–136 (2017)
6. Gössling, S., Scott, D., Hall, C.M.: Pandemics, tourism and global change: a rapid assessment of covid-19. Journal of sustainable tourism **29**(1), 1–20 (2020)
7. 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

8. Halpern, N., Graham, A.: Airport route development: A survey of current practice. Tourism Management **46**, 213–221 (2015). https://doi.org/https://doi.org/10.1016/j.tourman.2014.06.011, https://www.sciencedirect.com/science/article/pii/S0261517714001137
9. 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)
10. Huse, C., Evangelho, F.: Investigating business traveller heterogeneity: Low-cost vs full-service airline users? Transportation Research Part E: Logistics and Transportation Review **43**(3), 259–268 (2007)
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. Linden, E.: Pandemics and environmental shocks: What aviation managers should learn from covid-19 for long-term planning. Journal of Air Transport Management **90**, 101944 (2021)
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. Møller, A., Pera, A., Dalsgaard, J., Aiello, L.: The parrot dilemma: Human-labeled vs. llm-augmented data in classification tasks. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 179–192 (2024)
18. Nguyen, T.D.: An approach to improve the accuracy of rating prediction for recommender systems. Automatika **65**(1), 58–72 (2024)
19. Prabhakar, E., Santhosh, M., Krishnan, A.H., Kumar, T., Sudhakar, R.: Sentiment analysis of us airline twitter data using new adaboost approach. International Journal of Engineering Research & Technology (IJERT) **7**(1), 1–6 (2019)
20. Rizve, M.N., Duarte, K., Rawat, Y.S., Shah, M.: In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. arXiv preprint arXiv:2101.06329 (2021)
21. Schwenker, B., Wulf, T.: Scenario-based strategic planning: Developing strategies in an uncertain world. Springer Science & Business Media (2013)
22. Sobieralski, J.B.: Covid-19 and airline employment: Insights from historical uncertainty shocks to the industry. Transportation Research Interdisciplinary Perspectives **5**, 100123 (2020)
23. staff, T.: Chatgpt vs claude 3 test: Can anthropic beat openai's superstar? *Tech.co (2024)*
24. *Suau-Sanchez, P., Voltes-Dorta, A., Cugueró-Escofet, N.: An early assessment of the impact of covid-19 on air transport: Just another crisis or the end of aviation as we know it? Journal of Transport Geography* ***86****, 102749 (2020)*

25. *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)*
26. *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)*
27. *Sun, X., Wandelt, S., Zhang, A.: Covid-19 pandemic and air transportation: Summary of recent research, policy consideration and future research directions. Transportation research interdisciplinary perspectives* ***16****, 100718 (2022)*
28. *Wang, C., Nulty, P., Lillis, D.: Using pseudo-labelled data for zero-shot text classification. In: International Conference on Applications of Natural Language to Information Systems. pp. 35–46. Springer (2022)*
29. *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)*
30. *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`
31. *Wong, C.W., Cheung, T.K.Y., Zhang, A.: A connectivity-based methodology for new air route identification. Transportation Research Part A: Policy and Practice* ***173****, 103715 (2023)*
32. *Wu, S., Gao, Y.: Machine learning approach to analyze the sentiment of airline passengers' tweets. Transportation Research Record* ***2678****(2), 48–56 (2024)*
33. *Yang, W., Zhang, R., Chen, J., Wang, L., Kim, J.: Prototype-guided pseudo labeling for semi-supervised text classification. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 16369–16382 (2023)*
34. *Zahraee, S.M., Shiwakoti, N., Jiang, H., Qi, Z., He, Y., Guo, T., Li, Y.: A study on airlines' responses and customer satisfaction during the covid-19 pandemic. International Journal of Transportation Science and Technology* ***12****(4), 1017–1037 (2023)*
35. *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)*