Exploring Ryanair Passenger Feedback: Topic Insights and Customer Satisfaction Classification

Niccolò Settimelli

A.A. 2023/2024

Contents

1 Introduction

The low-cost airline market in Europe is increasingly competitive and complex, with Ryanair maintaining its position as a market leader despite the influx of new significant competitors. Recent developments, however, have led to a surge in negative feedback and criticisms directed towards Ryanair, resulting in a markedly damaged corporate image (*bbc article Ryanair disease*). An intriguing statistic highlights the stagnation in customer satisfaction indices over the past four years, reflecting a lack of substantial improvement (*yougov article low increase of perception score*) .

This project aims to address this issue not through a traditional sentiment analysis approach, which categorizes comments as either positive or negative, but by identifying and analyzing the key topics discussed by Ryanair's customers. By focusing on these topics, the project seeks to predict the sentiment of comments and provide actionable insights for Ryanair to enhance its customer satisfaction indices.

To achieve this, we will first preprocess the dataset, emphasizing the customer comments. We will then employ Latent Dirichlet Allocation (LDA) to uncover latent topics within the comments. Each topic will be assigned a relevance score using a Naive Bayes method to evaluate customer sentiment towards the topic. Subsequently, a logistic regression classifier will be developed to analyze how each topic influences overall customer satisfaction.

Furthermore, we will explore the application of these topics through pattern mining techniques to provide additional insights. This approach aims to offer Ryanair a comprehensive understanding of customer concerns and preferences, thereby facilitating targeted improvements in service and customer relations.

2 Dataset

The dataset utilized for this study is preprocessed and available on Kaggle (link to dataset). However, for the objectives of our analysis, an additional preprocessing phase is required. Initially, certain attributes present in the dataset are deemed irrelevant for both the preliminary text mining phase and for the subsequent analysis of topic-related sentiment. Consequently, these attributes are excluded from our analysis. Specifically, attributes such as user ratings for Cabin Staff Service, Food & Beverages, Ground Service, Value for Money, as well as comment IDs and Aircraft information, are removed.

After the removal of these attributes, the remaining dataset comprises only categorical attributes like 'Type of Traveler' and 'Seat Type', which describe the passenger's flight experience, along with the 'Comment', which is divided into 'Comment Title' and 'Comment'.

The dataset consists of 2,250 reviews collected from 2012 to 2024. Each review contains an average of 130 words and exhibits a relatively balanced distribution of sentiments, with 40% classified as positive and 60% as negative.

3 Text Mining Analysis

In this segment of the project, we will focus on the primary objective inspired by the study *Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews* link to paper. We will adopt similar methodologies throughout the project, with adjustments made as necessary due to the distinct characteristics of our dataset.

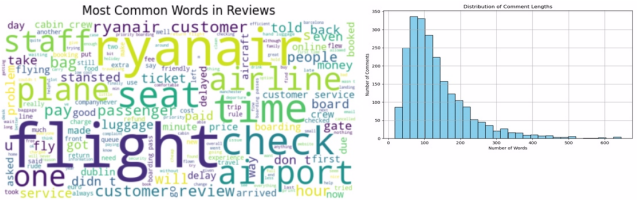

Figure 1: Comments visualization

3.1 Preprocessing Phase

During the preprocessing phase, we will exclusively utilize the features from the dataset, specifically the `Comment Title` and `Comment` fields, by combining them. We will then split the full comments into two distinct datasets: one designated for training the model and the other for testing.

For both datasets, we will apply preprocessing steps using the `CountVectorizer`, which includes:

- **Tokenization**, excluding non-English words
- **Stop-word Filtering** and **Stemming**

Subsequently, an additional preprocessing step is necessary for our specific objectives. Given that Latent Dirichlet Allocation (LDA) operates under the assumption that a document is a mixture of topics, it is beneficial to focus solely on nouns. Conversely, for sentiment analysis, which will be employed to weight each topic, we will use a Bayesian approach to assess adjectives, as they are fundamental to the sentiment expressed by users in their comments.

Before transforming the token vectors for both adjectives and nouns, the vector designated for LDA is further processed. LDA requires a matrix of word frequencies to be as dense as possible to effectively identify topics. Consequently, frequently occurring words on the same topic have been consolidated under a unified key.

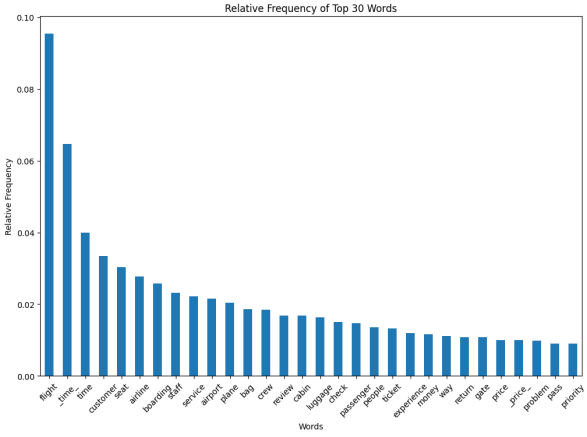

Figure 2: top words distribution in the Bow arrays

The Bag-of-Words (BoW) array has been constructed for all vectors, ensuring that each comment serves as an instance and the features represent the occurrences of each word. We opted against using Term Frequency-Inverse Document Frequency (TF-IDF) because, as specified in the referenced paper, LDA favors a simpler frequency-based approach.

3.2 LDA phase

The goal of applying Latent Dirichlet Allocation (LDA) in this study is to identify dimensions of customer satisfaction for Ryanair. LDA posits that each document (review) is represented as a probabilistic distribution over latent topics (θ), with a Dirichlet prior $\text{Dir}(\alpha)$. Each topic k has a distribution over words (ϕ_k) with a Dirichlet prior $\text{Dir}(\beta)$. The generative process involves: (1) choosing the number of words $N \sim \text{Poisson}(\xi)$ for a document, (2) selecting a topic distribution $\theta \sim \text{Dir}(\alpha)$, and (3) for each word w_n, choosing a topic $z_n \sim \text{Multinomial}(\theta)$ and then choosing the word w_n from $p(w_n \mid z_n, \beta)$. The key inferential task is to maximize the posterior distribution of the hidden variables (θ and z) given the observed words (w), formally:

$$p(\theta, z \mid w, \alpha, \beta) = \frac{p(\theta, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)}$$

where exact inference is computationally intractable. Thus, LDA requires as input a matrix of word frequencies per review and the number of topics to be extracted. One of the major challenges in Latent Dirichlet Allocation (LDA) is determining the number of topics without prior knowledge. As suggested by the paper, perplexity is used to find a sufficient number of topics for the model. The perplexity metric is defined as:

$$\text{Perplexity} = \exp\left(-\frac{\sum_{d=1}^{D} \log p(w_d \mid \theta_d, \phi)}{\sum_{d=1}^{D} N_d}\right)$$

where $p(w_d \mid \theta_d, \phi)$ is the probability of observing words w_d given topic proportions θ_d and word distributions ϕ, and N_d is the number of words in document d.

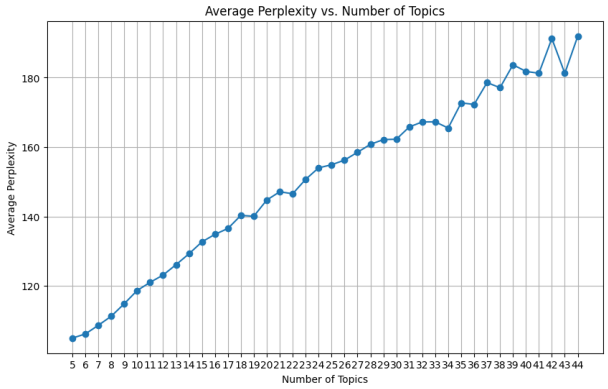

Figure 3: variation of perplexity per topic

A too-small number of topics results in a model that is too coarse and inaccurate, while too many topics can lead to a complex and less interpretable model. The paper with a larger dataset and pre-grouped comments studied the decrease in perplexity to determine the optimal number of topics. However, this approach was not feasible for our dataset due to the limited number of comments, which did not allow for effective use of perplexity. Instead, we followed an alternative approach discussed in another research (link to research), which suggests evaluating topic quality by examining the points where the perplexity graph flattens, while adhering to a threshold, in this case, 150. Based on this approach and considering interpretability, 19 topics were chosen for our model. Each of the 19 topics was reinterpreted using the top words from each topic's distribution in the reviews, and their probabilistic distribution across all comments was analyzed.

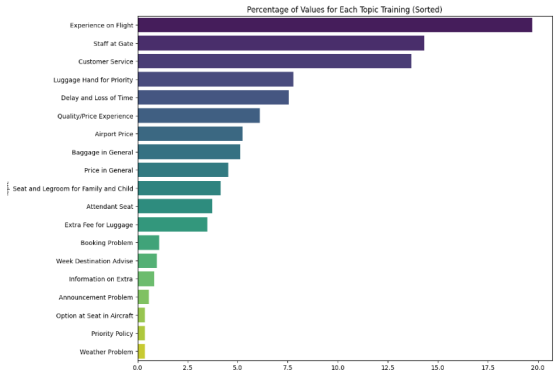

Figure 4: topic distribution probability

3.3 Adjective customer sentiment weight by Bayesian approach

For sentiment analysis, the matrix containing adjective frequencies was utilized to obtain weights using a Naive Bayesian approach. The conditional probability for each adjective given a class (e.g., recommendation or not) was computed as follows:

$$P(\text{Class} \mid \text{Adjective}) = \frac{P(\text{Adjective} \mid \text{Class}) \cdot P(\text{Class})}{P(\text{Adjective})}$$

This formula provides the probability of a class given the presence of an adjective, helping to assess how the adjective influences the likelihood of recommendation in a review. These weights were then normalized to fall within the range of -1 to 1 using the following formula:

$$\text{Normalized Weight} = \frac{X - \min(X)}{\max(X) - \min(X)} \times 2 - 1$$

This normalization allows for the evaluation of adjectives with either positive or negative impacts on the review. Only adjectives with high weights (above 0.75 or below -0.75) were considered to register significant positive or negative influences on the topic.

For each review, the sentiment score was calculated as follows:

$$\text{sentiment\_score} = \text{adj\_freq} \times \text{weight\_adj}$$

This sentiment score was then redistributed to each topic based on its probability of presence in the review.

3.4 Classification

With each review now containing 19 topics and their respective weights reflecting the positivity or negativity of the review towards those topics, the model

was evaluated to determine how these attributes influence the correct prediction of review sentiment. Additionally, the model was assessed to check if the predictions were excessively negative when compared to predictions made solely based on text using a Bayesian approach.

3.5 Logistic Regression Analysis

Logistic Regression was employed as the classification model for two primary reasons. First, it is particularly effective with numerical features, such as our topics. Second, it provides insights into how each topic influences the outcome relative to others.

To ensure the model's robustness, feature selection was performed to assess whether some attributes were redundant. This was achieved through the following steps:

- **Correlation Matrix:** Initially, a correlation matrix was used to identify any highly correlated features. Features that were excessively correlated were considered for elimination to avoid redundancy.

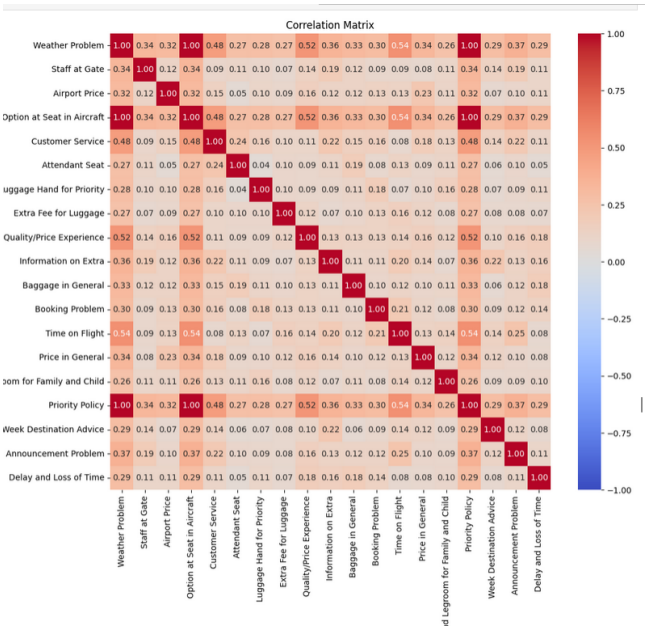

Figure 5: correlation matrix

- **Sequential Feature Selection:** Given the low inter-dependence among topics, Sequential Feature Selection was employed. This technique demonstrated that removing certain features had minimal impact on model accuracy. A plot of accuracy versus the number of features selected revealed

that accuracy remained stable at approximately 80% with up to 8 features. Consequently, the model was refined to include only these 8 features.

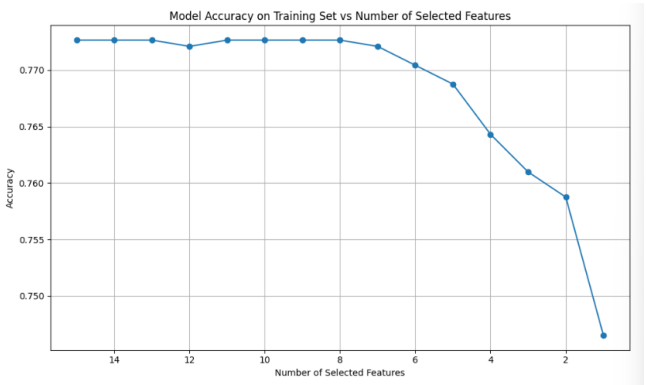

Figure 6: accuracy variation for removing features

- **Statistical Significance:** Post feature selection, p-values for each variable in the Logistic Regression model were found to be very low. This indicates that all selected features are significant for maintaining the model's accuracy.
- **Recall and Precision:** The model exhibited a higher recall relative to its precision. A high recall is particularly advantageous in identifying positive instances, which suggests that the model is effective at detecting relevant cases.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1826	0.068	-17.325	0.000	-1.316	-1.049
Staff at Gate	1.2080	0.341	3.539	0.000	0.539	1.877
Airport Price	-1.3474	0.798	-1.689	0.091	-2.911	0.216
Customer Service	2.1149	0.396	5.338	0.000	1.338	2.891
Luggage Hand for Priority	3.3816	0.616	5.490	0.000	2.174	4.589
Quality/Price Experience	2.3033	0.682	3.378	0.001	0.967	3.640
Time on Flight	4.9388	0.359	13.758	0.000	4.235	5.642
Seat and Legroom for Family and Child	3.9888	1.034	3.856	0.000	1.961	6.016
Week Destination Advice	2.1945	2.607	0.842	0.400	-2.915	7.304

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.92	0.86	287
1	0.80	0.60	0.69	163
accuracy			0.80	450
macro avg	0.80	0.76	0.77	450

Figure 7: Regression Results

Despite these promising results, it is crucial to compare the performance of the Logistic Regression model with that of the Naive Bayes classifier, which was previously used to obtain the weights. This comparison will help to determine whether the trade-off in accuracy is justified by the gain in interpretability of the topics.

3.6 Model Comparison

We conducted a comparison of the two models using two pipelines to analyze their general performance and ROC curves based on the test dataset. The results indicate:

```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.92      0.85       287
           1       0.80      0.58      0.67       163

    accuracy                           0.80       450
   macro avg       0.80      0.75      0.76       450
weighted avg       0.80      0.80      0.79       450

Naive Bayes Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.92      0.91       287
           1       0.86      0.83      0.84       163

    accuracy                           0.89       450
   macro avg       0.88      0.88      0.88       450
weighted avg       0.89      0.89      0.89       450

ROC AUC Logistic Regression: 0.84
ROC AUC Naive Bayes: 0.93
```

- **Performance Metrics:** There is a general decline in almost all performance metrics when transitioning from the Naive Bayes model to the Logistic Regression model. However, the reduction is not so severe as to negate the benefits of having features that explain the positivity or negativity of the reviews.

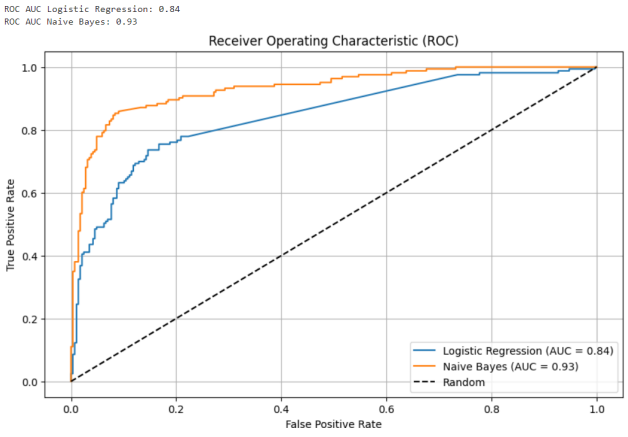

Figure 8: ROC curves comparison

- **Cross-Validation:** To quantify the reduction in accuracy, k-fold cross-validation was applied. This method provided a more comprehensive view of model performance by examining the consistency of results across different subsets of the data.
- **Residual Analysis:** The residuals from both models were compared. While the residual distribution was non-Gaussian, the Wilcoxon test confirmed that there is indeed a performance loss when switching from the Naive Bayes model to the Logistic Regression model.

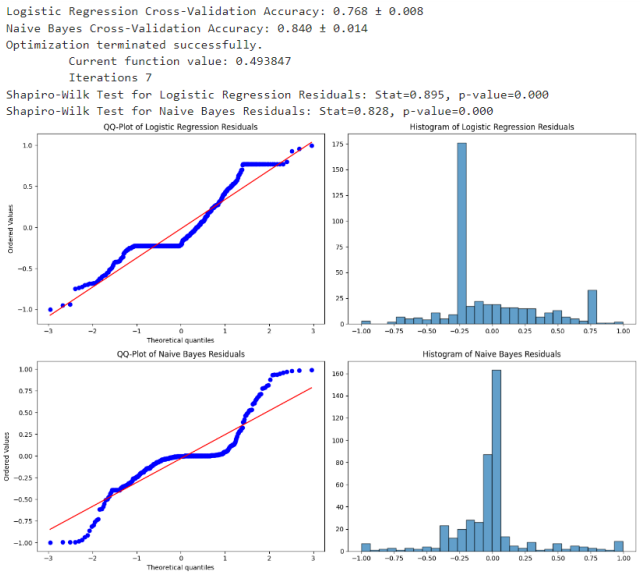

Figure 9: residual analysis

In summary, while the Logistic Regression model demonstrates good performance, especially with high recall, it shows a slight decrease in accuracy compared to the Naive Bayes model. This trade-off between accuracy and interpretability should be considered in the context of the specific application and the importance of feature interpretability.

4 Application Example

In this phase, we demonstrate how the identified topics can be utilized alongside additional customer travel information from the reviews. Specifically, we use frequent pattern mining to associate topics with travel characteristics through association rules.

4.1 Preprocessing and Data Visualization

To begin, we optimize the dataset by removing irrelevant features, such as outdated rankings, and eliminating features with excessive missing values, for instance, "Wifi Entertainment," which has more than half of its instances marked as NaN. We then illustrate how the topics represent the extent to which a review discusses a particular topic. This distribution is typically sparse, with many reviews discussing a topic minimally or not at all, resulting in scores frequently below 0.1 or above -0.1. To address this sparsity, we discretize the topics into three categories:

- **Good**: If the score for a topic is greater than 0.1.
- **Not Good**: If the score is less than -0.1.
- **Not Interesting**: For all other cases.

We apply this discretization and observe a high frequency of the "Not Interesting" category compared to the other two labels. Subsequently, we use statistical data visualization to analyze how topics impact company performance. For instance, we present a chart displaying the average monthly frequency of "Good" and "Not Good" ratings over a 12-year period for four topics: "Staff at Gate," "Time on Flight," "Quality/Price Experience," and "Customer Service." This visualization helps reveal seasonal trends in positive or negative topic sentiments throughout the year.

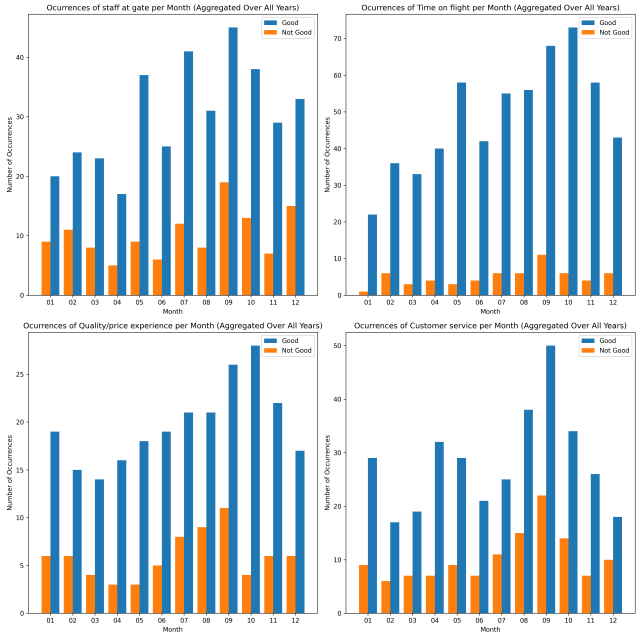

Figure 10: application example with seasonality of some topics

4.2 Pattern Mining Example

In addition to the topics, we consider the following features: 'Type Of Traveller', 'Origin', 'Destination', 'Passenger Country', 'Seat Type', and 'Month'. By unique labeling of the topics, where 'Good' and 'Not Good' labels are combined with the feature names, we apply the Apriori algorithm to identify frequent patterns excluding the "Not Interesting" category. Given that 'Good' or 'Not Good' labels are rare compared to 'Not Interesting', rare pattern mining with a low support threshold is necessary.

Once interesting patterns are identified, we focus on those with high confidence and also employ more complex metrics to validate them:

- **Lift**:
$$\text{Lift}(X, Y) = \frac{p(X, Y)}{p(X) \cdot p(Y)}$$
Lift measures how much more likely the occurrence of Y is given X, compared to its baseline probability.

- **Zhang's Metric**:
$$\text{Zhang Metric} = \frac{\text{conf}(X \to Y) - \text{conf}(\neg X \to Y)}{\max(\text{conf}(X \to Y), \text{conf}(\neg X \to Y))}$$
Zhang's Metric evaluates the strength of association between X and Y relative to X's baseline probability.

- **Conviction**:
$$\text{Conviction}(X, Y) = \frac{1 - p(Y)}{1 - p(X, Y)}$$
Conviction assesses the degree to which X and Y are dependent, with values greater than 1 indicating a strong association.

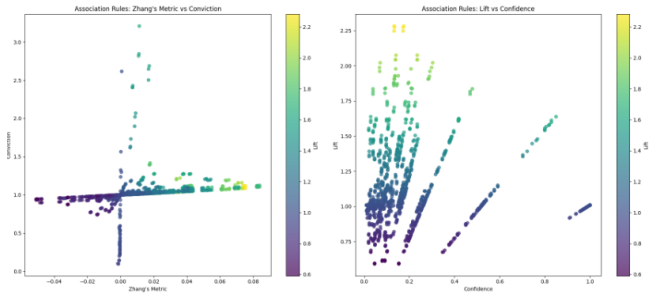

Figure 11: frequent pattern scatter plots

We plot these metrics and identify those with the best trade-off between confidence and complex metrics, such as with confidence ¿ 0.75 and lift ¿ 1.6.

For example, we found that travelers from Stansted Airport in the UK generally perceive a better quality-price ratio, and that negative comments in Economy Class are not typically related to service quality.

5 Graph Interface

In this work, models were downloaded in `pkl` format, and a prediction function was created to process input comments. The function starts by applying preprocessing steps to the comments, including vectorization as detailed previously. It then uses Latent Dirichlet Allocation (LDA) and Bayesian models to assign topics and weights to the processed input. The graphical interface, as illustrated in the figure, employs logistic regression to determine whether the comment is positive or negative. It then presents the topics from the logistic regression model sorted by their p-values in descending order. Topics with lower p-values are positioned at the top, as they indicate a higher significance in influencing the predicted outcome. A lower p-value in logistic regression suggests that the corresponding topic has a stronger impact on the prediction, thus providing a clearer indication of its relevance in determining the sentiment of the comment.

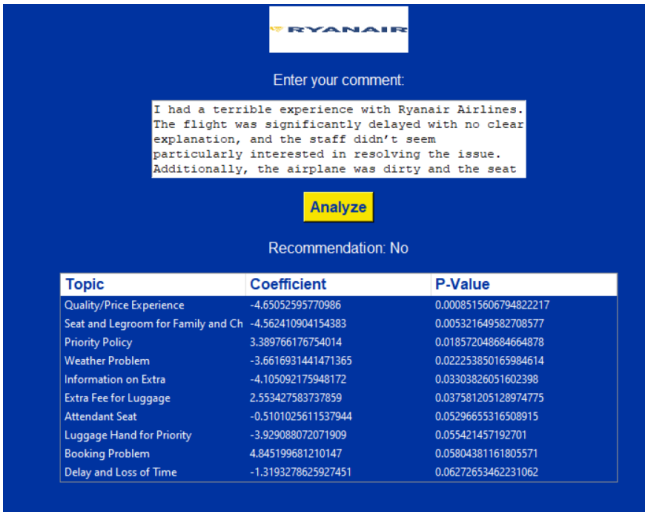

Figure 12: Interface example

6 Conclusion

In conclusion, the analysis of Ryanair passenger feedback provides valuable insights into customer satisfaction by leveraging a topic-based approach. The model developed, which employs Latent Dirichlet Allocation (LDA) to identify latent topics and logistic regression to classify overall satisfaction, achieves a commendable accuracy of 77% in predicting the sentiment of customer reviews.

Although this approach does not match the precision of traditional Bayesian models, it offers significant advantages by exploring new dimensions of customer satisfaction. By analyzing specific topics discussed in the reviews, Ryanair can identify critical areas and opportunities for improvement that might be overlooked by conventional sentiment analysis methods.

The application of pattern mining to these topics further enriches the analysis, revealing associations between flight characteristics and customer sentiment. For instance, it uncovers trends such as variations in satisfaction based on airport of departure or differences in perceived value-for-money across service classes.

This multidimensional approach not only enhances the understanding of customer feedback but also provides actionable insights for refining service strategies and marketing efforts. Implementing these insights can lead to informed decision-making, optimized operations, and ultimately, improved corporate image.

In summary, this work demonstrates the value of moving beyond simple sentiment classification to explore the nuanced aspects of customer satisfaction. The integration of topic modeling and pattern mining techniques not only deepens the analysis of customer feedback but also paves the way for continuous innovation and improvement in customer service management at Ryanair.