Understanding Diverging Numerical and Text-based Sentiment by Segment in Airline Reviews

Group 1: Geethan Sundaram, Stephanie Diau, Alison Miller

DS4002

FEBRUARY 16, 2026

Project Details

- **Motivation**
  - The sentiment of a text may not truly be representative of its true meaning...
    - **Ex: “Oh man, I sure love this increase in prices!” → Sentiment: ☺️ | True meaning: ☹️**
  - Understand if any other label can cause this mismatch between sentiments of airline reviews and the true rating associated with the review
- **Hypothesis**
  - Alternative: at least ONE category significantly alters the probability of mismatch between the review content and the overall rating.
- **Research Question**
  - What proportion of the airline reviews are “mismatched”, and are these “mismatch” cases associated with the type of traveler or cabin flown?
- **Modeling Approach**
  - Skytrax User Reviews Dataset → Sentiment Analysis and Mismatch Classification → Logistic Regression
- **GOAL:** Examine cases where the positivity or negativity of a user’s airline review and that user’s numeric overall rating for the airline appear to disagree

Data Explanation/Acquisition

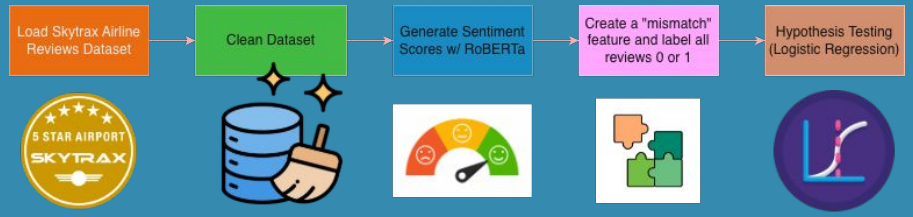
From Skytrax Text Based Data:

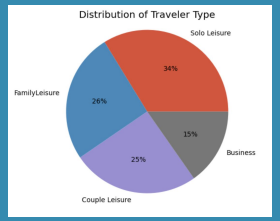
Data Dictionary

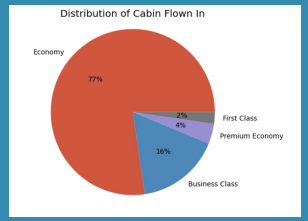
Column	Description
content	The text content of each review left by the reviewer. This is the text that the sentiment analysis would be run on.
overall_rating	A numerical integer rating of the airline given by the reviewer, from 1–10 inclusive.
type_traveller	A category to describe the type of traveller the reviewer is (or in other words, the reviewer's reason to travel).
cabin_flown	The cabin flown by the reviewer.

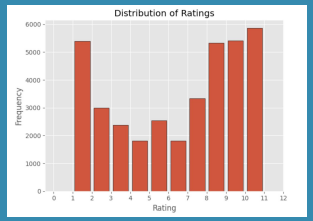
- **Acquisition:** Data set found on Github, content compiled in data set from *Skytrax*.
- **Ethical or Licensing Concerns:** No concerns, Creative Commons Zero v1.0 Universal in license.
- **Size:** 41,396 responses to 36,861 responses, 20 columns to 4 columns.
  - Preserved only relevant columns
  - Eliminated rows missing content, overall_rating.
- **Necessary Techniques:**
  - From cleaning: normalizing whitespace, punctuation fixes, unfilled responses for categorical variables to *'Unknown'*.
  - Migrated categorical variables to dummy variable in separate .csv(s).

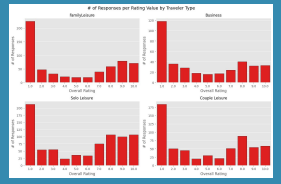
Analysis Plan and Justification

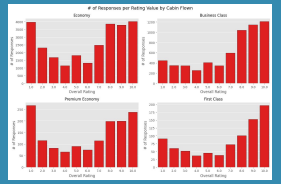


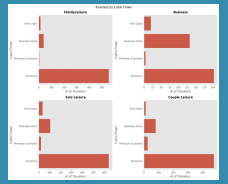


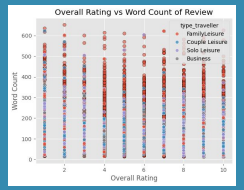


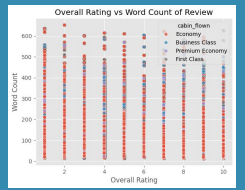












- We started off cleaning the data
  - Remove missing text reviews, replace missing categories with “Unknown”, clean text
- We did EDA, helping us learn the spread and composition of the data
- Then we performed sentiment analysis on all reviews (RoBERTa) and then classified all reviews as “mismatched” or not
- We then used logistic regression
  - H0: No category is significant
  - H1: At least ONE category is significant
- Then we evaluated our results
  - Interpretation of p-values
  - Odds ratios
  - Performance metrics

Tricky Analysis Decision

- This is **foreshadowing**, but we did see that based on our logistic regression model, there did not seem to be evidence of significance in any categories for influencing review-rating mismatch
- We initially were not sure how to take this conclusion and thought if we should try another model or a different p-value for the existing model with fear our analysis was flawed
- We ultimately decided to stick with the results we obtained since they were representative of our initial hypothesis and thoughtfully selected analysis and evaluation plan

Bias and Uncertainty Validation

- In terms of the reviews, we were aware that some bias would exist. We tried to standardize all the reviews during data preprocessing to ensure the main reason for differences in sentiments is the text content rather than external factors (whitespace, grammar, null values).
- Other potential bias
    - Voluntary response bias
    - Class imbalance
- Uncertainty
    - Std errors and p-values
    - Odds ratios w/ confidence intervals
    - Confusion matrix and R^2

Results/Conclusion

- Though we technically reject the null hypothesis because cabinflown_Unknown is significant (p-value < 0.05), cabinflown_Unknown represents missing values which means that it does not provide any insight on the effect of our two variables.
- Our odds ratios are close to 1, indicating minimal impact on the chances of a mismatch and our confusion matrix predicted 0 mismatches using our test data, which might point to class imbalance
- Additionally, our model fit was extremely weak at ~0.023, indicating low predictive power.
- This supports our final conclusion that despite our technical hypothesis interpretation, our results show that neither variable has a substantial effect on the likelihood of a mismatch.

```
Call:
glm(formula = mismatch ~ type_traveller + cabin_flown, family = "binomial",
    data = traindata)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.97799    0.28686 -10.381   <2e-16 ***
type_travellerCouple Leisure  0.27340    0.34503   0.792    0.428
type_travellerFamilyLeisure   0.31781    0.34147   0.931    0.352
type_travellerSolo Leisure    0.21055    0.33220   0.634    0.526
type_travellerUnknown         0.25480    0.28552   0.892    0.372
cabin_flownEconomy         -0.08508    0.06658  -1.278    0.201
cabin_flownFirst Class     -0.14980    0.17915  -0.836    0.403
cabin_flownPremium Economy -0.02959    0.13771  -0.215    0.830
cabin_flownUnknown          1.30540    0.08687  15.028   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14202  on 29489  degrees of freedom
Residual deviance: 13874  on 29481  degrees of freedom
AIC: 13892

Number of Fisher Scoring iterations: 5
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6892  479
         1    0    0

               Accuracy : 0.935
                 95% CI : (0.9291, 0.9405)
    No Information Rate : 0.935
    P-Value [Acc > NIR] : 0.5122
```

	OddsRatio	2.5 %	97.5 %
(Intercept)	0.05089488	0.02756388	0.08565026
type_travellerCouple Leisure	1.31442169	0.68147057	2.66649575
type_travellerFamilyLeisure	1.37411635	0.71854771	2.77179421
type_travellerSolo Leisure	1.23436213	0.65973760	2.45298420
type_travellerUnknown	1.29020027	0.76895629	2.37696449
cabin_flownEconomy	0.91844222	0.80724400	1.04805555
cabin_flownFirst Class	0.86088361	0.59673012	1.20681297
cabin_flownPremium Economy	0.97084559	0.73592991	1.26366494
cabin_flownUnknown	3.68917696	3.11177862	4.37468546

Next Steps

Different Categories

Airline name, author country, aircraft, route

More Specific

Specifically WiFi? Food? Entertainment?

Different Model/Approach

Better Statistical Model: Interaction terms, Additional params, Multilevel

Predictive Accuracy: Random Forests Gradient Boosting, Decision Tree

References

- [1] Pongsathon Pookduang, Rapeepat Klangbunrueang, Wirapong Chansanam, and Tassanee Lunrasri, “Advancing Sentiment Analysis: Evaluating RoBERTa against Traditional and Deep Learning Models,” *Engineering Technology & Applied Science Research*, vol. 15, no. 1, pp. 20167–20174, Feb. 2025, doi: https://doi.org/10.48084/etasr.9703.
- [2] quankiquanki, “GitHub - quankiquanki/skytrax-reviews-dataset: An air travel dataset consisting of user reviews from Skytrax (www.airlinequality.com),” GitHub, 2025. https://github.com/quankiquanki/skytrax-reviews-dataset

Github: https://github.com/gthsun/DS-4002-Project-1

Q & A ?