IT2386 TEXT AND SOCIAL ANALYTICS PROJECT

PROJECT PROPOSAL

Submitted By:	Gangula Karthik (223715Y)*
	Ng Jun Ming (220080D)
	Seah Pin Shien (220273H)
Team Name:	Dip Sum
Tutor:	Ms Jane Zhang Mr Lim Sing Tat
Submission:	14 January 2024

Table of Contents

Introduction

Word of mouth (WOM) plays a significant role in influencing and forming consumer's choices. 92% of consumers are more likely to believe the recommendations from friends and family over all other forms of advertising [1] and it is said to drive a significant 6 trillion dollars' worth of annual consumer spending [2]. In today's highly competitive airline market, maintaining long-lasting customer relationships becomes the key factor to business success. Identifying the factors influencing word of mouth and customer loyalty can give these airlines an edge in the market and drive sustained growth.

With the ongoing digital transformation, online communication and virtual interactions have become very common nowadays, which has led to the growth of electronic word of mouth (eWOM) / online WOM and now the online reviews as user generated content (UGC) have become a very large part of WOM. This is especially true for intangible products and services (airlines, hotels, and other tourism products), as it is difficult to try out the products before consumption, as a result a lot of reviews turn to online reviews to learn about the reputation of the airline. In Singapore, 30% of the survey respondents highlighted that they would look at reviews before choosing an airline, which was third highest number among the other countries, with UAE and Hong Kong, placing first and second respectively [3].

Given the importance of these reviews, a text classification model will be developed for this project with the aim being to classify whether a review is coming from a promoter, detractor, or passives (neutral). This will allow airlines to refine their WOM marketing strategies to influence more promoters for their airlines, as a result build up the customer loyalty and sustain long term growth.

Background information

Consumers increasingly utilize online tools to share their opinions and discover more about the products and services that they consume [4]. eWOM, unlike its predecessor, is not limited to face-to-face interactions. It has a much broader and lasting impact. This is true with especially negative reviews, where 96% of customers specifically look for negative reviews [5]. However, the benefits of a positive review are significant as well. Harvard Business Review found that a one-star increase in a Yelp rating can lead to a 5-9% increase in business revenue [6]. Looking at the airline industry specifically, online reviews were heavily relied upon by potential air passengers, especially when assessing factors like safety, price, and service quality. During the COVID-19 pandemic, the focus had shifted towards safety measures, with negative reviews related to safety perceptions having a significant impact on the decision-making process of passengers [7]. Another example is Qatar airlines, which capitalizes on WOM marketing to boost sales [8]. Based on the examples given, it can be understood that the dynamics of eWOM in the airline industry is essential for developing strategies that enhance customer experience, build brand loyalty, and ultimately drive business success.

Business scenario

The primary audience comprises of the **Airline Strategic Planning and Marketing Teams**, these professionals are tasked with interpreting data-driven insights to enhance the customer experience, refine marketing approaches, and develop comprehensive strategies for business growth. They would utilize the information provided by the predictive model to make informed decisions that convert detractors to promoters, retain loyalty of promoters, and address the concerns of the passives to improve overall business performance.

Objectives identified

The primary objective is to develop a classification model that excels in performance based on key metrics, allowing airlines to leverage this tool to gain insights and enhance customer experiences derived from their reviews. This classification model will be able to accurately determine whether a review about the airline is indicative of a promoter, detractor, or passive.

To achieve this the team is in its commitment to perform:

a. Data visualization and insights: Utilize advanced data visualization tools to analyze patterns and trends in airline customer reviews across multiple data sources. This will assist in pinpointing influential factors, refining the data preprocessing steps, and crafting enhanced features that lay the groundwork for the development of a robust classification model.

b. Extensive Experimentation with different models: Utilize various machine learning classification algorithms such as Decision Trees, Random Forests, Logistic Regression and more to categorize stocks as bullish or bearish based on the analysis of earnings call transcripts.

c. Performance Benchmarks: The F1 score is a more robust measure compared to accuracy, especially when handling a dataset with substantial implications for false positives and negatives, such as airline reviews. Misclassifying a promoter as a detractor, or vice versa, could lead airlines to misjudge their service quality and market position. A high F1 score is crucial as it ensures the model accurately identifies true promoters and detractors, which is critical for airlines to make informed decisions that affect customer satisfaction and loyalty. The aim is to refine the model to achieve a high F1 score, balancing precision and recall, and avoiding bias towards any particular class.

Delegation of tasks/responsibility

Websites that each member is to scrape data from:

1. Pin Shien: https://www.airlineratings.com/airline-passenger-reviews/

2. Jun Ming: https://www.airlinequality.com/

3. Karthik: https://www.tripadvisor.com/Airlines

Tasks Each member is responsible for:

Name	Individual Tasks
Karthik	• Help to oversee the project • Communicate with relevant stakeholders • Work on the Introduction, background, and business scenario
Jun Ming	• Work on the timeline of the project • Determine tools and libraries to use for the project
Pin Shien	• Work on business objectives of project • Help come up with project risks

Below the project timeline has been provided for the 3 different phases of the project which is the project initialization phase, data collection and preparation phase, and finally the modelling and evaluation phase.

Project Initialization Phase

Name	Assignee	Status	Priority	Start date	Due date	Projec...	Project Progress
Project Initialization and Business Understanding 7	NM K SS	IN PR...		12/13/23	Tomorrow	Planning	71%
Project Initialization and Business Understanding Form team and assign roles.	K NM SS	COM...		12/13/23	12/15/23	Planning	100%
Project Initialization and Business Understanding Discuss and agree on project goals.	SS K NM	COM...		12/18/23	12/22/23	Planning	100%
Project Initialization and Business Understanding Research on problem statements and topics	NM K SS	COM...		12/19/23	12/25/23	Planning	100%
Project Initialization and Business Understanding Research and identify potential data sources	K NM SS	COM...		12/18/23	12/25/23	Planning	100%
Project Initialization and Business Understanding Draft the project proposal	K NM SS	COM...		12/26/23	6 days ago	Planning	100%
Project Initialization and Business Understanding Discuss with Stakeholders	NM SS K	IN PR...		3 days ago	2 days ago	Planning	0%
Project Initialization and Business Understanding Review and finalize the proposal for submission	K NM SS	IN PR...		3 days ago	Tomorrow	Planning	0%

Figure 1.1 – tasks for project initialization phase

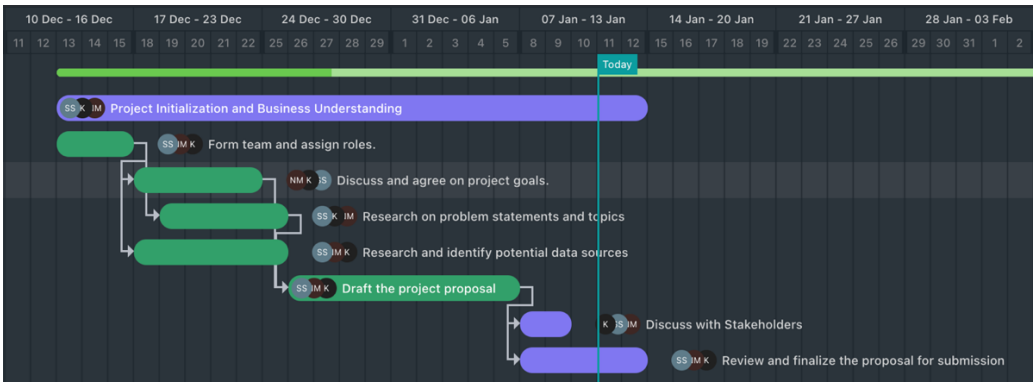

Figure 1.2 – Gantt chart for project initialization

Data Collection and Preparation Phase

Data collection and Preparation 5	K NM SS	TO DO		Mon	Jan 29	Execution	0%
Data collection and Preparation web scraping to collect data	NM K SS	TO DO		Mon	Jan 18	Execution	0%
Data collection and Preparation data understanding	K NM	TO DO		Jan 19	Jan 22	Execution	0%
Data collection and Preparation data preprocessing	K NM SS	TO DO		Jan 23	Jan 26	Execution	0%
Data collection and Preparation Combine data and resolve inconsistencies	NM SS K	TO DO		Jan 25	Jan 27	Execution	0%
Data collection and Preparation submit data to stakeholder	K NM SS	TO DO		Jan 29	Jan 29	Execution	0%

Figure 2.1 – Tasks for data collection and preparation phase

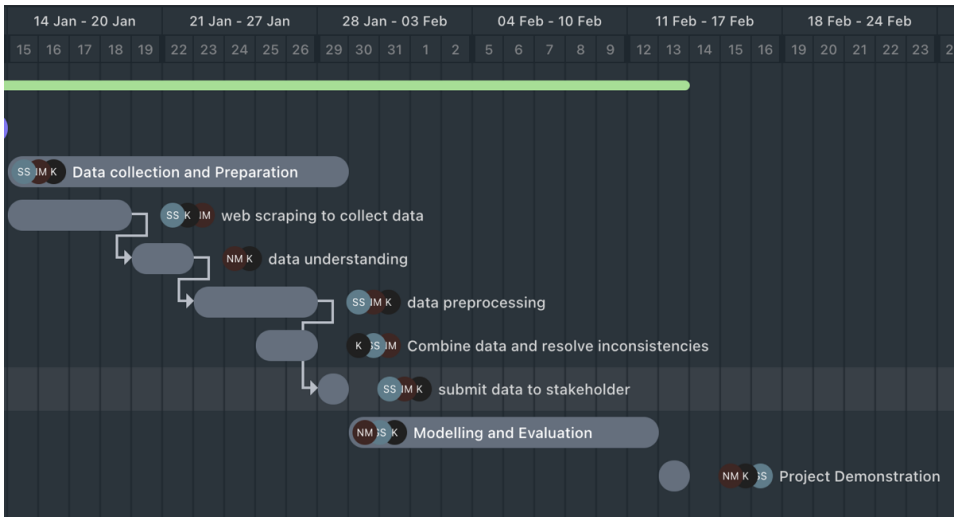

Figure 2.2 – Gantt chart for data collection and preparation phase

Modelling and Evaluation Phase

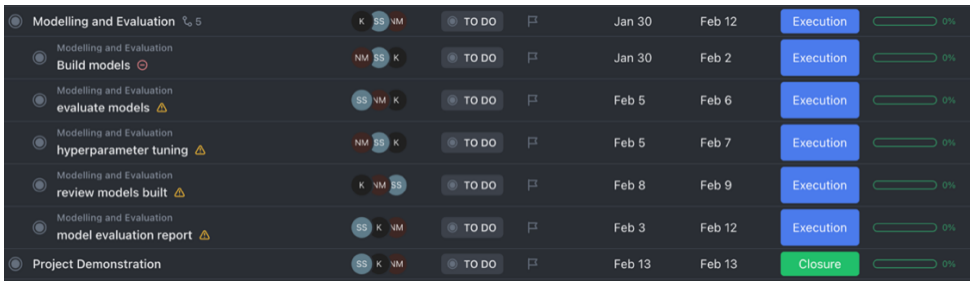

Figure 3.1 – Tasks for modelling and evaluation

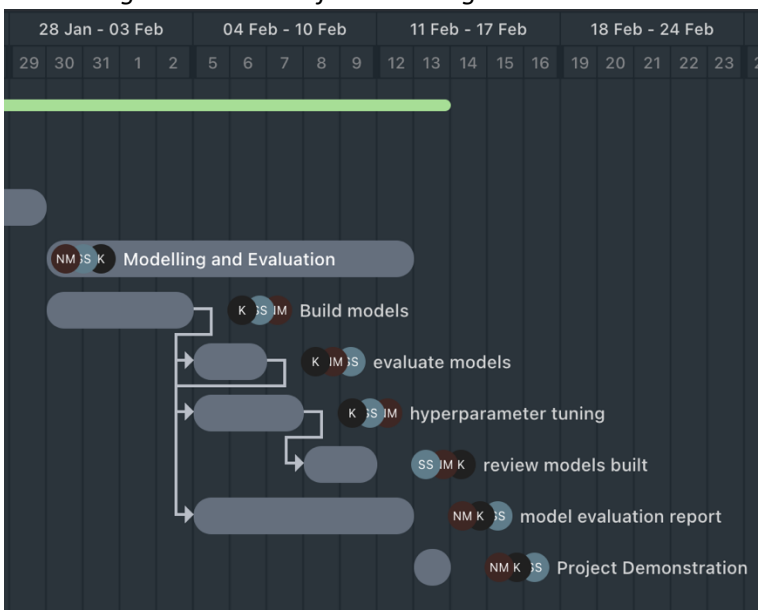

Figure 3.2 – Gantt chart for modelling and evaluation

Clickup Chart Link: https://app.clickup.com/9016101548/v/li/901600787286

Tools used

Data Gathering:

- Python
- Selenium (library for web scraping)
- Beautiful Soup (library for parsing HTML and XML)
- Yfinance library (API call to yahoo finance to get stock data)

Data Pre-processing and Understanding:

- NLTK (Natural Language Toolkit)
- NumPy and Pandas (Python library for data manipulation and analysis)
- Matplotlib (library for data visualization)
- Seaborn (statistical data visualization based on Matplotlib)

Data Preparation and Balancing:

- scikit-learn (sklearn) (machine learning library for data preprocessing and modeling)
- imbalanced-learn (imblearn) (library for oversampling/under sampling)

Modeling and Hyperparameter Tuning:

Tools and libraries:

- scikit-learn (sklearn) (machine learning library for model construction and training)
- Hyperparameter tuning using scikit-learn tools.

Project Management:

- Project Management software: Click Up
- Code Collaboration Software: GitHub

Potential Risks

Data Collection and Validation:

Web scraping will be utilized to acquire the necessary data. When using the chosen approach, it's essential to acknowledge the potential risk associated with anti-scraping measures implemented by the target website. In the event of encountering such obstacles, the supervisor will be promptly informed, and an alternative website will be used.

Model Development and Prototyping:

The data scraped will be used to train the machine learning model. Use a portion of the data for validation to prevent overfitting. Plan for peer reviews and incorporate feedback for model improvement.

Project Documentation and Backups:

To ensure that the work done is not lost or deleted version control systems like git will be used to track changes and backup work regularly.

PDPA concerns:

There may be some ethical concerns about PDPA, thus we will strictly only look at the review text and not things like the name and date posted. Thus, is to make sure that the identity of the reviewer is undisclosed.

Data Sources

During web scraping we will be getting the data from 3 different websites:

1. https://www.airlineratings.com/airline-passenger-reviews/
2. https://www.tripadvisor.com.sg/
3. https://www.airlinequality.com/

- **Data Source:** Reviews and ratings from a specific website.
- **Textual Data (Independent Variable):** Reviews/articles about airlines and other features derived from feature engineering on the text.
- **Target Data (Dependent Variable):** Derived from overall rating scores.
    - **5-Star Rating Adaptation:** If a 5-star rating system (like TripAdvisor) is used, a specific mapping (Figure 4.1) will be applied to convert these ratings to the NPS scale.
    - **10-Star Rating Interpretation:**
        - Scores of 8 and above: Active promoters.
        - Scores of 6 and below: Detractors.
        - Scores between 6 and 8: Neutral.
- **Target Column Classification:** Three classes - promoters, passives, detractors (Figure 4.2).
    - Promoters: Reviews with scores of 8 and above.
    - Detractors: Reviews with scores of 6 and below.
    - Passives: Reviews with scores between 6 and 8.

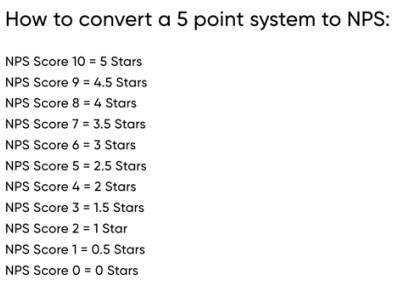

Fig 4.1 – convert 5 stars to NPS

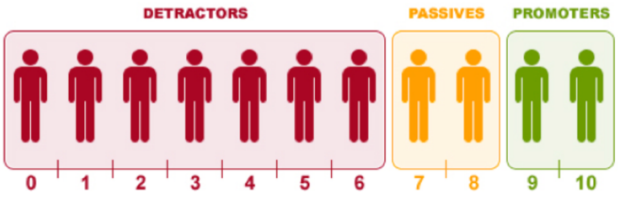

Fig 4.2 – classify into promoters, detractors and passives

Appendix

1. https://www.forbes.com/sites/kimberlywhitler/2014/07/17/why-word-of-mouth-marketing-is-the-most-important-social-media/?sh=64de9d2f54a8

2. https://www.lxahub.com/stories/word-of-mouth-marketing-stats-and-trends-for-2023

3. https://business.yougov.com/content/45395-global-how-much-do-social-media-reviews-matter-air

4. https://www.sciencedirect.com/science/article/abs/pii/S0148296309001969

5. https://www.pinmeto.com/blog/electronic-word-of-mouth-ewom#:~:text=Unlike%20traditional%20word%20of%20mouth,reviews%20when%20they%20shop%20online.

6. https://www.hbs.edu/ris/Publication%20Files/12-016_a7e4a5a2-03f9-490d-b093-8f951238dba2.pdf

7. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.983987/full

8. https://www.referralcandy.com/blog/qatar-airways-marketing-strategy

9. https://www.simplr.ai/blog/stars-smileys-and-thumbs-how-to-measure-customer-service-success