TBA2105 Web Mining

Revised Project Plan

Predicting SIA Stock Trends Using Airline Industry Sentiment Analysis

Student:	Kelvin Chong Kean Siong
Group:	6

Course:	TBA2105 Web Mining
Environment:	R (Windows), RStudio
Document Date:	October 26, 2025

Executive Summary

This project predicts **Singapore Airlines (SIA)** stock price movements by analyzing sentiment from **airline industry-wide news**. By collecting broader industry coverage rather than only SIA-specific articles, we ensure sufficient data volume for robust sentiment analysis. The approach is validated on 4-5 comparison carriers to demonstrate generalizability, but **SIA remains the primary analytical focus** throughout.

Key Revision	Rationale
Industry-wide data collection	Ensures 10-50 articles/day vs. 1-3/week for SIA-only
SIA remains primary focus	80% of analysis centers on SIA stock prediction
4-5 validation tickers	Proves method generalizability (20% of effort)
Google News + Yahoo RSS	Stable, ToS-compliant sources

Table of Contents

1. Objectives & Scope

2. Data Sources & Collection Strategy

3. Data Acquisition & Storage

4. Text Processing & Sentiment Analysis

5. Feature Engineering

6. Modeling Strategy

7. Evaluation

8. Validation on Other Tickers

9. Exploratory Data Analysis & Visualizations

10. Risk Mitigation Strategies

11. Success Criteria Summary

12. Ethical Considerations

13. Key References

14. Quick Start Guide

1. Objectives & Scope

(CRISP-DM: Business Understanding)

Primary Goal

Predict **next-day price direction for SIA stock** (C6L.SI) by integrating:

• Airline **industry-wide news sentiment** (broader data collection)

• Historical SIA stock prices and technical indicators

• Macroeconomic factors affecting aviation (oil prices, travel indices)

Why Industry-Wide Data Collection?

Problem	SIA-specific news articles are sparse (1-3 per week)
Solution	Collect airline industry news affecting all carriers (fuel costs, travel demand, regulations)
Rationale	Industry trends impact SIA directly → richer sentiment signal with sufficient daily volume (10-50 articles/day)

Coverage

Primary Analysis Target:

• **SIA** (C6L.SI) - Singapore Airlines

Validation Tickers (4-5 carriers for method validation):

• **Cathay Pacific** (0293.HK) - Regional competitor

• **Delta Airlines** (DAL) - Major US carrier

• **ANA** (9202.T) - Major Asian carrier

• **Lufthansa** (LHA.DE) - Major European carrier

• *Optional 5th:* **Qantas** (QAN.AX) or **Air France-KLM** (AF.PA)

Role of Validation Tickers: Prove that industry sentiment → stock prediction works systematically, not just by chance for SIA. These will receive **minimal individual analysis** (summary statistics only).

Prediction Framework

Horizon: Daily (today's industry sentiment → tomorrow's SIA closing direction)

Classification: 3-class {UP, DOWN, FLAT}

• UP: Next-day return > +0.2%

• DOWN: Next-day return < -0.2%

• FLAT: Within ±0.2%

Alternative: Binary {UP vs NOT-UP} for trading interpretation

Success Criteria

Primary: Macro-F1 score ≥ 0.55 for SIA prediction (beating naive baseline)

Validation: Method achieves F1 > 0.50 on at least 3/5 tickers (proves generalizability)

Insight: Identify which news topics (fuel, demand, safety) most impact SIA

2. Data Sources & Collection Strategy

(CRISP-DM: Data Understanding)

2.1 News Data - INDUSTRY-WIDE FOCUS

Philosophy: Cast a wide net to capture all airline industry signals

Primary News Sources (Industry Coverage)

Source Type	Examples
Global Aviation News	Reuters Aviation/Transport, Bloomberg Airlines & Aerospace, FlightGlobal, Yahoo Finance, Simple
Search Keywords	"airline industry", "aviation sector", "air travel", "fuel costs", "pilot strike", "Boeing", "Airbus"
Singapore/Asia Specific	Straits Times Business, Channel NewsAsia (CNA), Business Times, South China Morning Post

Why NOT Only SIA-Specific?

Approach	Data Volume	Result
SIA-only	~1-3 articles/week	Data sparsity → unreliable daily sentiment
Industry-wide	10-50 articles/day	Sufficient volume → robust sentiment features

2.2 Stock Market Data

Primary: SIA (C6L.SI)

• Source: Yahoo Finance via {quantmod}/{tidyquant}

• Fields: Adjusted close, open, high, low, volume

• Period: 2020-01-01 to present (~5 years for train/test)

Validation Tickers: Cathay, Delta, ANA, Lufthansa (same fields, same period)

Exogenous Market Variables:

• **Brent Crude Oil** (BZ=F) - Major cost driver for airlines

• **USD Index** (DX-Y.NYB) - Currency impact on international carriers

• **VIX** (^VIX) - Market volatility proxy

• **Singapore STI** (^STI) - Local market sentiment

2.3 Data Volume Targets

Data Type	Target Volume	Timeframe
News Articles	8,000-10,000 articles (avg ~15/day)	2 years
Stock Data (SIA)	~1,200 trading days	2020-2024
Usable Observations	~1,000 after feature lags	After cleaning

3. Data Acquisition & Storage

(CRISP-DM: Data Preparation - Part A)

3.1 R Environment Setup

Required Packages:

Core data manipulation: tidyverse, lubridate, arrow

Web scraping: rvest, httr2, xml2, polite

Financial data: quantmod, tidyquant

Text processing: tidytext, textdata, SnowballC, tm

Modeling: tidymodels, themis, xgboost, ranger, glmnet

Evaluation & visualization: yardstick, timetk, patchwork

3.2 Project Directory Structure

```
TBA2105_SIA_Prediction/
├── README.md
├── renv.lock
├── .gitignore
├── data_raw/ # Raw scraped data (not in Git)
│ ├── news_html/
│ └── scrape_logs/
├── data_interim/ # Cleaned data
│ ├── news_clean.parquet
│ ├── prices_sia.parquet
│ └── macro_vars.parquet
├── data_features/ # Model-ready datasets
├── R/ # Analysis scripts
│ ├── 01_scrape_news.R
│ ├── 02_scrape_prices.R
│ ├── 03_clean_text.R
│ ├── 04_sentiment_analysis.R
│ ├── 05_feature_engineering.R
│ ├── 06_modeling_sia.R
│ ├── 07_validation_tickers.R
│ └── 08_evaluation.R
├── reports/
├── figs/
└── models/
```

3.3 News Scraping Implementation

Recommended News Sources (Updated):

Tier	Source	Type	Reliability
1	Google News RSS (aviation keywords)	RSS Feed	High - Most reliable
1	Yahoo Finance Business RSS	RSS Feed	High - Stable
2	FlightGlobal RSS	RSS Feed	Medium - May require access
3	Reuters (via Google News)	Aggregated	High - Indirect access

3	Straits Times Business	HTML Scraping	Medium - Anti-scraping measures

Key Implementation Features:

• Polite scraping with rate limiting (1 request/second)

• Robust error handling with logging to data_raw/scrape_logs/

• Keyword filtering for airline-related content

• Deduplication based on headlines

• Storage in Parquet format for efficient I/O

4. Text Processing & Sentiment Analysis

(CRISP-DM: Data Preparation - Part B)

4.1 Text Cleaning Pipeline

Step	Operation	Purpose
1	Normalization	Lowercase, remove URLs, strip whitespace
2	Tokenization	Unigrams + bigrams (e.g., "fuel cost")
3	Cleaning	Remove stopwords, punctuation, numbers
4	Stemming	Porter stemmer via {SnowballC}
5	Filtering	Keep tokens with document frequency ≥ 3

4.2 Sentiment Scoring (Multi-Tier Approach)

Tier 1: Lexicon-Based Sentiment

Use established sentiment dictionaries to score words, then aggregate by ticker and date:

• **Bing lexicon:** Binary positive/negative classification

• **AFINN:** Valence scoring (-5 to +5 scale)

• **NRC:** Emotion categories (positive, negative, fear, trust)

Daily Aggregate Features (per ticker-date):

• sent_score: Net sentiment (-1 to +1)

• sent_share_pos: Proportion of positive words

• sent_article_count: Volume of coverage

• sent_mean, sent_median, sent_sd: Statistical aggregates

Tier 2: Financial-Specific Lexicon (Optional)

• Loughran-McDonald financial dictionary

• Custom airline keywords: "profit warning", "guidance upgrade", "load factor decline"

Tier 3: Transformer-Based (Stretch Goal)

• FinBERT via {reticulate} for headline scoring → P(positive), P(negative), P(neutral)

5. Feature Engineering

(CRISP-DM: Data Preparation - Part C)

5.1 Technical Indicators for SIA

Category	Features
Returns	ret_lag1, ret_lag2, ret_lag5 (1, 2, 5 days)
Volatility	20-day rolling standard deviation, 5-day volatility
Momentum	RSI(14), MACD, Bollinger Bands
Volume	Volume z-score (vs 30-day average), Volume ratio
Moving Averages	MA(5), MA(20), MA(50), momentum ratios

5.2 Macro Variables Processing

• **Oil price changes:** Brent 1-day return, 5-day return, 20-day MA

• **USD strength:** USD Index 1-day return

• **Market volatility:** VIX level, VIX change

• **Singapore market:** STI 1-day return, 5-day return

Critical: All macro features lagged by 1 day to avoid lookahead bias.

5.3 Master Feature Matrix for SIA

Join all features on (ticker, date) key:

• Price features from SIA stock data

• Sentiment features from daily news aggregation

• Macro features from exogenous variables

• Fill missing sentiment with neutral (0)

• Create label: ret_next → {UP, DOWN, FLAT}

Temporal Alignment: Only news published before market close on date T contributes to T's features.

6. Modeling Strategy

(CRISP-DM: Modeling)

6.1 Problem Framing

• **Multi-class classification:** Predict {UP, DOWN, FLAT}

• **Binary sensitivity check:** UP vs. NOT-UP for trading simulation

6.2 Algorithm Suite

Model	Purpose	Rationale
Logistic Regression (multinomial)	Baseline	Interpretable coefficients, fast
Random Forest	Nonlinear interactions	Robust, feature importance
XGBoost	Performance champion	Handles imbalance, regularization
Naive Bayes	Text-heavy variant	Original proposal inclusion

6.3 Time-Series Cross-Validation

Rolling origin with expanding window:

• 400-day initial training window

• 60-day validation window

• Skip 30 days between folds

• Cumulative (expanding window)

• **No data leakage:** Features at time T use only information available ≤ T

6.4 Handling Class Imbalance

Compare strategies:

• **Class weights:** Penalize minority class errors more

• **SMOTE:** Synthetic oversampling of minority classes

• **Down-sampling:** Random undersampling of majority class

7. Evaluation

(CRISP-DM: Evaluation)

7.1 Primary Metric

Macro-F1 Score: Average F1 across all 3 classes (handles imbalance)

7.2 Secondary Metrics

- **Accuracy:** Overall correct predictions
- **Per-class Precision & Recall:** Identify which direction is hardest to predict
- **Confusion Matrix:** Systematic bias patterns (e.g., over-predicting FLAT)
- **Matthews Correlation Coefficient (MCC):** Balanced measure for imbalanced sets

7.3 Baseline Comparisons

Baseline	Description	Expected F1
Random guess	33.3% accuracy for 3-class	~0.33
Always-FLAT	Predict majority class	~0.35-0.40
Sentiment-only	Remove price features	TBD
Price-only	Remove sentiment features	TBD

8. Validation on Other Tickers

8.1 Purpose

Apply the SIA-trained model to validation tickers (Cathay, Delta, ANA, Lufthansa) to prove the method generalizes across carriers, not just SIA specifically.

8.2 Process

1. Build feature matrices for each validation ticker (same pipeline as SIA)

2. Apply finalized SIA model without retraining

3. Calculate Macro-F1 for each ticker

4. Report summary statistics only (no deep-dive analysis)

8.3 Success Criterion

Method achieves **F1 > 0.50 on at least 3 out of 5 tickers** $\rightarrow$ proves systematic predictive power

9. Exploratory Data Analysis & Visualizations

9.1 Visualization Plan

Sentiment-Price Correlation:

Time-series plot with dual-axis (price + sentiment), lag correlation analysis

Industry Dynamics:

Correlation heatmap of returns across all airline tickers, oil price vs. airline basket

Text Analysis:

Word clouds (positive/negative terms), bigram network graphs, topic timeline (if using LDA)

Feature Importance:

SHAP values for tree models (top 20 features), permutation importance

SIA Spotlight:

Singapore-specific news sources contribution, event case studies

10. Risk Mitigation Strategies

Risk	Likelihood	Impact	Mitigation
Insufficient news volume	Medium	High	Collect from 5+ diverse sources; use 4+ years data
Scraper breaks	High	Medium	Prioritize RSS feeds; robust error handling
Severe class imbalance	Medium	Medium	Widen thresholds (±0.5%); use SMOTE
Poor model performance	Medium	High	Ablation studies; feature selection; ensemble methods
No generalization	Low	High	Focus on SIA-only story; emphasize method exploration
Copyright/ToS violations	Low	Critical	Store only headlines; cite sources; respect robots.txt

Contingency Plan:

If Macro-F1 < 0.50 by Week 6: (1) Simplify to binary (UP vs. NOT-UP), (2) Focus on specific events (earnings, fuel spikes), (3) Pivot to descriptive analysis ("What news topics correlate with SIA volatility?")

11. Success Criteria Summary

Must-Have (Core Requirements)

- Collect ≥5,000 airline industry news articles
- Build daily sentiment features for SIA
- Train 3+ models (Logistic, RF, XGBoost)
- Achieve SIA Macro-F1 ≥ 0.55 on test set
- Demonstrate method on 4-5 validation tickers
- Produce final report + presentation

Nice-to-Have (Stretch Goals)

- FinBERT transformer-based sentiment
- Topic modeling (LDA) for thematic analysis
- Intraday prediction (if minute-level data available)
- Economic simulation (hypothetical trading returns)
- Interactive Shiny dashboard for exploration

12. Ethical Considerations

12.1 Data Collection Ethics

- **Respect for Content Creators:** Only scrape publicly available content; attribute sources
- **Server Load:** Rate limiting (1 req/sec) to avoid overwhelming servers
- **Robots.txt Compliance:** Automated checks before each scrape
- **Terms of Service:** Review and comply with each site's ToS

12.2 Model Use Ethics

- **No Financial Advice:** Clear disclaimer that models are for educational purposes only
- **Transparency:** Disclose limitations (historical bias, lookahead bias risks)
- **Bias Awareness:** Acknowledge that news sentiment may reflect media biases
- **Privacy:** No personal data collected (only public news articles)

12.3 Academic Integrity

- **Citation:** All sources (code libraries, papers, data providers) properly cited
- **Original Work:** All analysis and code written by student
- **Reproducibility:** Full code + data (or data collection scripts) provided

13. Key References

13.1 Academic Papers

• Tetlock, P. C. (2007). "Giving content to investor sentiment: The role of media in the stock market." *Journal of Finance*, 62(3), 1139-1168.

• Bollen, J., Mao, H., & Zeng, X. (2011). "Twitter mood predicts the stock market." *Journal of Computational Science*, 2(1), 1-8.

• Loughran, T., & McDonald, B. (2011). "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks." *Journal of Finance*, 66(1), 35-65.

13.2 R Packages Documentation

• **Tidyverse:** https://www.tidyverse.org/

• **Tidymodels:** https://www.tidymodels.org/

• **Tidyquant:** https://business-science.github.io/tidyquant/

• **Tidytext:** https://www.tidytextmining.com/

• **Rvest:** https://rvest.tidyverse.org/

13.3 Data Sources

• **Yahoo Finance API:** https://finance.yahoo.com/

• **Google News RSS:** https://news.google.com/

• **IATA Economics:** https://www.iata.org/en/publications/economics/

13.4 Sentiment Lexicons

• **Bing Liu Lexicon:** https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

• **AFINN:** http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

• **Loughran-McDonald:** https://sraf.nd.edu/loughranmcdonald-master-dictionary/

14. Quick Start Guide

After Setup (Week 1 Complete)

Step 1: Install Packages

```
renv::restore()
```

Step 2: Run Data Collection (~2-3 hours)

```
source("R/01_scrape_news.R")
source("R/02_scrape_prices.R")
```

Step 3: Process Text & Calculate Sentiment (~30 minutes)

```
source("R/03_clean_text.R")
source("R/04_sentiment_analysis.R")
```

Step 4: Engineer Features (~15 minutes)

```
source("R/05_feature_engineering.R")
```

Step 5: Train Models (~2 hours with tuning)

```
source("R/06_modeling_sia.R")
```

Step 6: Validate & Evaluate (~30 minutes)

```
source("R/07_validation_tickers.R")
source("R/08_evaluation.R")
```

Step 7: Generate Reports

```
rmarkdown::render("reports/01_EDA.Rmd")
rmarkdown::render("reports/02_Model_Results.Rmd")
rmarkdown::render("reports/03_Final_Report.Rmd")
```

Summary of Key Changes from Original Plan

What Changed:

• Data Collection Scope: Industry-wide news (10-50 articles/day) instead of SIA-only (1-3/week)

• Primary Focus: SIA remains the main analytical target (80% of effort)

• Validation Purpose: 4-5 tickers used only to prove method generalizes (20% of effort)

• News Keywords: "airline industry", "aviation sector" vs. "Singapore Airlines" only

What Stayed the Same:

• SIA as primary subject of final report

• 3-class classification (UP/DOWN/FLAT)

• CRISP-DM methodology

• R environment & Tidymodels framework

• Deliverables (report, slides, code)

Key Message for Professor: "I expanded data collection to the airline industry as suggested, ensuring sufficient volume (~10,000 articles vs. ~150 SIA-only). However, my analysis remains centered on predicting **SIA stock trends**, using industry-wide sentiment as the input signal. The validation tickers (Cathay, Delta, ANA, Lufthansa) simply prove the method works systematically, not randomly. SIA is the star of the show—the others are supporting cast."