Leveraging News Sentiment to Forecast GDP's Demand Components: A Malaysian Case Study

Agenda

01 Project Background

02 Problem Statements

03 Aim, Questions & Objectives

04 Scope & Significance of Project

05 Literature Review

06 Methodology

07 Implementation

08 Results & Analysis

09 Discussion & Conclusions

Project Background

GDP & Economic Health:

GDP is crucial for assessing economic status, with increases indicating growth and decreases indicating decline.

Sentiment Analysis via Surveys:

Traditional surveys like BCI and CSI measure economic sentiments but face challenges in timeliness and representativeness.

Robust Sentiment Analysis Need:

The pandemic underscored the need for agile, technology-enhanced sentiment analysis methods due to survey limitations.

Technological Advances in Analysis:

Advancements in digital media and computational linguistics now enable efficient analysis of large text volumes for sentiment.

Economic Indicators' Importance:

Private investment and private consumption, along with imports and exports, are crucial sentiment-driven indicators for understanding economic dynamics, unlike government spending.

Problem Statements

Nowcasting vs. Publication Delays

- Economic indicators like GDP are often delayed, impacting timely analysis and decision-making. This issue is prevalent in both developing and developed countries, including the U.S. and the U.K., with Malaysia experiencing a 90-day delay in GDP data release.

Importance of Nowcasting

- The COVID-19 pandemic highlighted the critical need for nowcasting for real-time decision-making by governments and central banks to address economic disruptions .

Continuous Monitoring

- Despite the traditional focus on forecasting, nowcasting is essential for ongoing economic monitoring and proactive decision-making, even in stable times.

Problem Statements (cont'd)

Language Diversity in Sentiment Analysis

- Previous research on sentiment analysis from news articles has been predominantly in English. There is a gap in capturing sentiments from non-English sources, significant in multilingual contexts like Malaysia.

Nowcasting vs. Forecasting Clarification

- The study clarifies that while it uses "forecasting" due to quarterly data intervals, its core lies in applying high-frequency indicators (HFIs) for nowcasting within a forecasting framework, focusing on news sentiment to gauge economic conditions.

Aim

To assess the role of text-derived news sentiment from Mandarin-based news articles, in the forecasting of the Malaysian GDP and its four demand-side components of the expenditure approach namely, private investment, private consumption, and import and export, using machine learning techniques.

Project Questions & Objectives

Project Questions

1. Does the news sentiment index which is computed in the study accurately reflect the BCI and CSI published by the MIER?
2. How was the performance of the Mandarin text-derived news sentiment when compared to the English text-derived news sentiment in the work of Chong et al. (2021)?
3. Which of the four demand-side components (private investment, private consumption, imports and exports) of GDP exhibited a strong correlation with the newly constructed news sentiment index, and which of the components showed a weak correlation to the index?

Objectives

[1] To evaluate the ability of Mandarin-based news sentiment indices in predicting sentiments reflected by the BCI and CSI figures.

[2] To compare the performance of the Mandarin text-derived news sentiments to that of the English text-derived news sentiments reported in previous published work.

[3] To evaluate the correlation between the four demand-side components of GDP and the news sentiment index.

Scope of Project

Features of Dataset

- Dataset includes a selection from 8 online newspaper portals for news sentiment extraction: Asia Times, China Press, Guangmin Daily, Kwong Wah Jit Poh, Nanyang Siang Pau, Overseas Chinese Daily News, See Hua Daily News, and Sin Chew Jit Poh.
- Analysis capped at 10,000 articles across all portals.
- Macroeconomic indicators (GDP, private consumption, investment, imports, exports) sourced from Department of Statistics Malaysia (DOSM) at a quarterly basis.
- MIER survey-based indices (BCI and CSI) are sourced from the Malaysian Institute of Economic Research at a quarterly basis.

Lexicon and Dictionaries

- Mandarin lexicon by Jiang et al. (2019), referred to as JMZ, used for Mandarin news sentiment analysis.
- JMZ chosen for its comprehensive approach tailored to financial sector sentiments in Mandarin news.
- JMZ was based on the Loughran and McDonald (2011) framework and further enhanced through manual screening, and the word2vec algorithm expansion.
- Stop word dictionaries by Diaz & fseasy (2020) employed to eliminate common and irrelevant words.

Scope of Project (cont'd)

Geographical Region

- Study focuses on the Malaysian economy at the national level.

Domain of Study

- Analysis limited to topics covered by the dataset, excluding subjects outside this scope.

Time Frame

- Study's sample period: 2022 to 2023.
- Unlike Chong et al. (2021)'s broader timeframe (Q1 2006 - Q2 2021), this study is limited by time constraints, restricting its coverage.

Scope of Project (cont'd)

Software, Tools and Libraries

Software and Tools:

- ParseHub (v2.4.35) for scrapping news articles from the news portals chosen.
- Visual Studio Code (v1.91.1) for executing the remaining parts of the project.

Libraries:

- *pandas*, *numpy*, *statsmodels*, *unicodedata*, *jieba*, *re*, *json*, *matplotlib*, *scikit-learn* and *xgboost* libraries from Python and the respective submodules of *sklearn* and *statsmodels* for preprocessing, statistical tests, visualization, evaluation, and modelling.

Evaluation Metrics

Evaluation Metrics for Nowcasting Activity:

- Pearson correlation coefficients and statistical significance in linear regression to assess news sentiment's nowcasting ability for MIER indices.

Evaluation Metrics for Forecasting Activity:

- Pearson correlation coefficients between macroeconomics variables and the news sentiment index.
- Ratio RMSE and MAE compares non-baseline model's RMSE and MAE to those of the baseline model [OLS-AR(1)].
- RMSE and MAE ratios value < 1 indicate forecast improvement over baseline model.

Scope of Project (cont'd)

Predicting BCI & CSI

Linear Regression Model

- Linear Regression applying JMZ Mandarin Lexicon.

Baseline Model

- Ordinary Least Squares with Autoregressive order 1 [OLS-AR(1)].

Forecasting GDP and Components

- Ridge Regression Model
- LASSO Regression Model
- Random Forest
- Extreme Gradient Boosting
- Support Vector Machine

Significance of Project

01 Vernacular Language Sentiment Analysis

This study addresses the gap in sentiment analysis literature that lacks focus on vernacular languages, specifically Mandarin within the Malaysian context. It aims to deepen the understanding of economic sentiment in Malaysia, serving as a model for multiethnic and multilingual economies and enriching global economic sentiment analysis discussions.

02 High-Frequency Indicators (HFIs) for Forecasting

With limited research on the use of HFIs like news sentiment for forecasting macroeconomic variables in Malaysia, this study seeks to enhance local academic resources in this field. It emphasizes the importance of contextualizing methodologies and conclusions to fit Malaysia's unique economic, social, and political landscape, aiding central banks and policymakers in creating timely and accurate policies to foster economic growth and stability.

Literature Review :

Forecasting Key Elements of Interest through Text-Derived News Sentiments

Key Insights and Themes

- Convergence of economic forecasting and sentiment analysis has gained significant attention.
- Need for domain-specific sentiment analysis models for economic forecasting.
- Importance of exploring how different big data measures and textual data influence economic forecast accuracy.
- Significance of examining the predictive ability of sentiment indicators across different economic landscapes and times of crises.
- Interest in integrating sophisticated machine learning techniques with sentiment analysis for economic forecasting.
- Potential benefits of integrating sentiment analysis with other economic indicators for more robust predictive models.

Literature Review :

Forecasting Key Elements of Interest through Text-Derived News Sentiments (cont'd)

Gaps and Future Research Directions

- Development of domain-specific lexicons and sentiment-scoring models for economics.
- Comprehensive examination of a broader array of big data indicators and textual data beyond traditional news articles.
- Inclusion of sentiments expressed in multiple languages for representative and comprehensive analysis.
- Comparative studies across different economies to evaluate the generalizability and reliability of sentiment indicators.
- Exploration of complex machine learning models beyond conventional approaches for deeper insights.
- Integration of sentiment indicators with traditional economic data for a more holistic view of economic trends.

Studies Mentioned: Thorsrud (2016), Ashwin et al. (2021), Kalamara et al. (2022), Shapiro et al. (2022), Barbaglia et al. (2023), Barbaglia et al. (2024)

Literature Review :

Multilingual Sentiment Analysis with Emphasis on Economic and Financial Texts in Chinese

Key Insights and Themes

- Sentiment analysis has been applied to various languages, each presenting unique challenges and opportunities.
- Notable contributions to sentiment analysis from studies on multiple languages, including Algerian, Dravidian languages, Roman Urdu, Hindi, Malay, and various Arabic dialects.
- Chinese language sentiment analysis has focused on various applications, from social media to e-commerce platforms.
- There's a notable lack of research on sentiment analysis in Chinese financial and economic contexts.
- The call for more sophisticated deep learning methods and robust datasets across different linguistic domains.
- The unique challenges of code-mixing in sentiment analysis.

Literature Review :

Multilingual Sentiment Analysis with Emphasis on Economic and Financial Texts in Chinese (cont'd)

Gaps and Future Research Directions

- Need for resources, tools, and advanced NLP approaches specifically tailored to lesser-studied languages and dialects.
- Development of sentiment analysis models that are culturally sensitive and capable of real-time analysis for Chinese textual data.
- Expansion of sentiment analysis research to include financial and economic aspects within the Chinese language.
- The refinement of sentiment classification beyond binary categories, towards more nuanced understanding.

Studies Mentioned: Huang et al. (2020), Moudjari & Akli-Astouati (2020), Muhammad Fakhrur Razi Abu Bakar et al. (2020), Ong et al. (2020), Wang and Alfred (2020), Gupta et al. (2021), Abir Masmoudi et al. (2021), Rana et al. (2021), Al Shamsi & Sherief Abdallah (2022), Kogilavani Shanmugavadivel et al. (2022), Wei et al. (2022), Hegde et al. (2023), Tan & Zhang (2008), Zhang et al. (2018), Yang et al. (2020)

Literature Review :

Review of Machine Learning Models for Sentiment Analysis

Key Insights and Themes

- Machine learning techniques are increasingly applied to improve GDP forecasting and nowcasting.
- Various models have been tested across different economies, including developed and emerging markets.
- Ensemble methods and model combinations often outperform individual models.
- Tree-based methods (Random Forest, Gradient Boosting, XGBoost) consistently show strong performance.
- Regularized regression methods (Ridge, LASSO) are effective for high-dimensional datasets.
- Novel data sources (e.g., high-frequency indicators, news sentiment) can improve forecast accuracy.
- Model performance varies across countries, economic conditions, and forecast horizons.
- There's a trade-off between model complexity and interpretability.

Literature Review :

Review of Machine Learning Models for Sentiment Analysis (cont'd)

Gaps and Future Research Directions

- Need for further research on model selection based on specific economic contexts and data characteristics.
- Exploration of methods to handle mixed-frequency data and publication lags.
- Investigation of ways to improve model interpretability without sacrificing predictive power.
- Further research on incorporating alternative data sources into forecasting models.
- Addressing challenges related to data quality and availability, especially for emerging economies.
- Exploration of deep learning models (e.g., LSTM) with larger datasets.
- The choice of model for this study consists of LASSO, Ridge, XGBoost, Random Forest, and SVR models.

Studies Mentioned: Chong et al. (2021), Chu & Qureshi (2023), Zhang et al. (2023), Richardson et al. (2021), Muchisha et al. (2021), Nakazawa (2022), Dauphin et al. (2022), Agu et al. (2022), Kant et al. (2022), Martin (2019), Jasni et al. (2022)

Literature Review :

Chong et al. (2021) and News Sentiment Analysis for Macroeconomic Forecasting

Key Insights and Themes

- The study focuses on the impact of news sentiment on forecasting Malaysia's GDP and its components.
- Utilization of over 720 thousand daily news articles from English-based online portals in Malaysia, covering business and financial news from 2006 to June 2021.
- Employment of Loughran and McDonald (2011), and Correa et al. (2017) lexicons for sentiment scoring, and a model by Shapiro et al. (2022) for improved accuracy.
- Comprehensive text preprocessing including the removal of unnecessary elements and conversion to lowercase but excluding stemming and lemmatization due to the lexicons used.
- Monthly news sentiment indices were used for nowcasting survey-based sentiment measures (BCI and CSI).
- Forecasting was conducted using both linear and non-linear machine learning models, with news sentiment as the predictor variable.
- Evaluation of news sentiment's nowcasting and forecasting abilities through various statistical measures and machine learning model comparisons.

Literature Review :

Chong et al. (2021) and News Sentiment Analysis for Macroeconomic Forecasting (cont'd)

Gaps and Future Research Directions

- Suggestion to explore text-derived news sentiment from non-English based sources (Malay, Chinese, or Tamil) to reflect Malaysia's diverse society.
- Expansion of news sentiment analysis beyond the business and financial domains.
- The need for a more granular dataset and the application of non-linear machine learning models for refined forecast precision.

Studies Mentioned: Chong et al. (2021), Correa et al. (2017), Loughran & McDonald (2011), Shapiro et al. (2022), Nothman et al. (2019)

Project Methodology

- The proposed project adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology.
- CRISP-DM consists of six stages: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.
- The project will provide detailed discussions for each applicable stage, excluding the Deployment stage, as it falls outside the scope of this study.

Stage 1: Business Understanding

- The task of business understanding has been addressed in the previous slides.

Stage 2: Data Understanding

- 10,000 news articles from early 2022 to late 2023 will be scrapped using the features within ParseHub, from the selection among the 8 news portals.
- The dataset will be detailed in terms of its structure, content, and any initial observations about data quality of the news articles.
- Data on the five macroeconomic indicators (GDP, private investment, private consumption, imports, and exports) for the same period, on a quarterly basis, will be sourced from the official platform of the Department of Statistics Malaysia (OpenDOSM, 2024-b), and subjected to preliminary exploration.
- Data on the BCI and CSI, on a quarterly frequency, will be sourced from the Malaysian Institute of Economic Research for the same time period and will undergo preliminary data exploration.
- Exploration of all datasets will be conducted in Visual Studio Code and problematic entries will be identified and noted.

Stage 3: Data Preparation

Textual Dataset

Textual Data Preprocessing in Visual Studio Code:

- Cleaning: Missing entries will be removed, and incorrectly formatted entries will be corrected or eliminated.
- Segmentation: Articles will be segmented into individual Mandarin characters using the *jieba* library, then saved in a structured format for subsequent analysis.
- Stop words filtered out based on the list by Diaz & fseasy (2020).
- Punctuation, special characters, hyperlinks, HTML tags, and extra white spaces removed.
- Sentiment analysis conducted using the JMZ lexicon. Negations have already been accounted for in the JMZ lexicon.

Stage 3: Data Preparation (cont'd)

Textual Dataset

News Sentiment Index Computation:

- Articles organized by publication date.
- Sentiment score computed for each article as the net number of positive words minus negative words, relative to the total word count, then multiplied by 1000 to normalize per thousand words.
- Scores transformed into an index format; values above 100 indicate more positive sentiments, and below 100, more negative.
- Scores scaled by the total number of articles for each news portal and time period to create a news sentiment time series for each of the portals among the selection.
- Averaging the indices across all portals to form a single, aggregated news sentiment index, with a chosen frequency (quarterly).

$$sentiment\ index_k = 100 + \frac{\Sigma Positive_k - \Sigma Negative_k}{Total\ Word\ Count_k} \times 1000$$

Note: k represents the individual articles

Stage 3: Data Preparation (cont'd)

Numerical Dataset

Numerical Data Preprocessing:

- Normalization: Implement normalization procedures to ensure consistency across different data scales.
- Imputation: Apply imputation techniques for any missing values to maintain data completeness.
- Outlier Detection and Treatment: Identify and address outliers to prevent distortion in the analysis.

Stage 4: Modelling

Modelling Stage Overview:

- Divided into two parts: nowcasting BCI and CSI values using the news sentiment index; and forecasting five macroeconomic indicators (GDP, private investment, private consumption, imports, and exports) using machine learning models.

Nowcasting the BCI and CSI Values:

- The aim is to determine if the quarterly news sentiment index is an adequate indicator of the current quarter's BCI and CSI values.
- Nowcasting involves estimating the BCI and CSI for each quarter.

$$BCI_t = \alpha + \beta\, BCI_{t-1} + n\, s_t + \varepsilon_t$$

$$CSI_t = \alpha + \beta\, CSI_{t-1} + n\, s_t + \varepsilon_t$$

Stage 4: Modelling (cont'd)

Forecasting the Five Target Variables Using Machine Learning Models:

- Machine learning models will be used for forecasting the quarter-on-quarter changes in GDP, private investment, private consumption, imports, and exports.
- A rolling window approach will be adopted for model training.
- The forecasting process mimics real-world scenarios, using historical economic data and news sentiment for future economic outcome predictions.
- Training windows move forward by one quarter for each forecast, maintaining a fixed window size and sequentially generating forecasts with updated data.

Stage 5: Evaluation

Evaluation Stage Overview:

- Divided into two parts: the effectiveness of nowcasting BCI and CSI values using the news sentiment index; and assessing the accuracy of forecasting the five macroeconomic indicators with machine learning models.

Evaluation of the Nowcasting Activity

- **Pearson Correlation Coefficients:** This metric is used to evaluate the strength and direction of the linear relationship between the news sentiment index and the BCI and CSI values. A high and positive correlation coefficient indicates that the sentiment index effectively mirrors the fluctuations in BCI and CSI values, showcasing its adequacy as an indicator.

- **Statistical Significance in Linear Regression:** The significance value of the sentiment index in linear regression analysis assesses its predictive power for BCI and CSI values. A statistically significant sentiment index underscores its relevance and reliability in nowcasting BCI and CSI values.

Stage 5: Evaluation (cont'd)

Evaluation of the Forecasting Activity

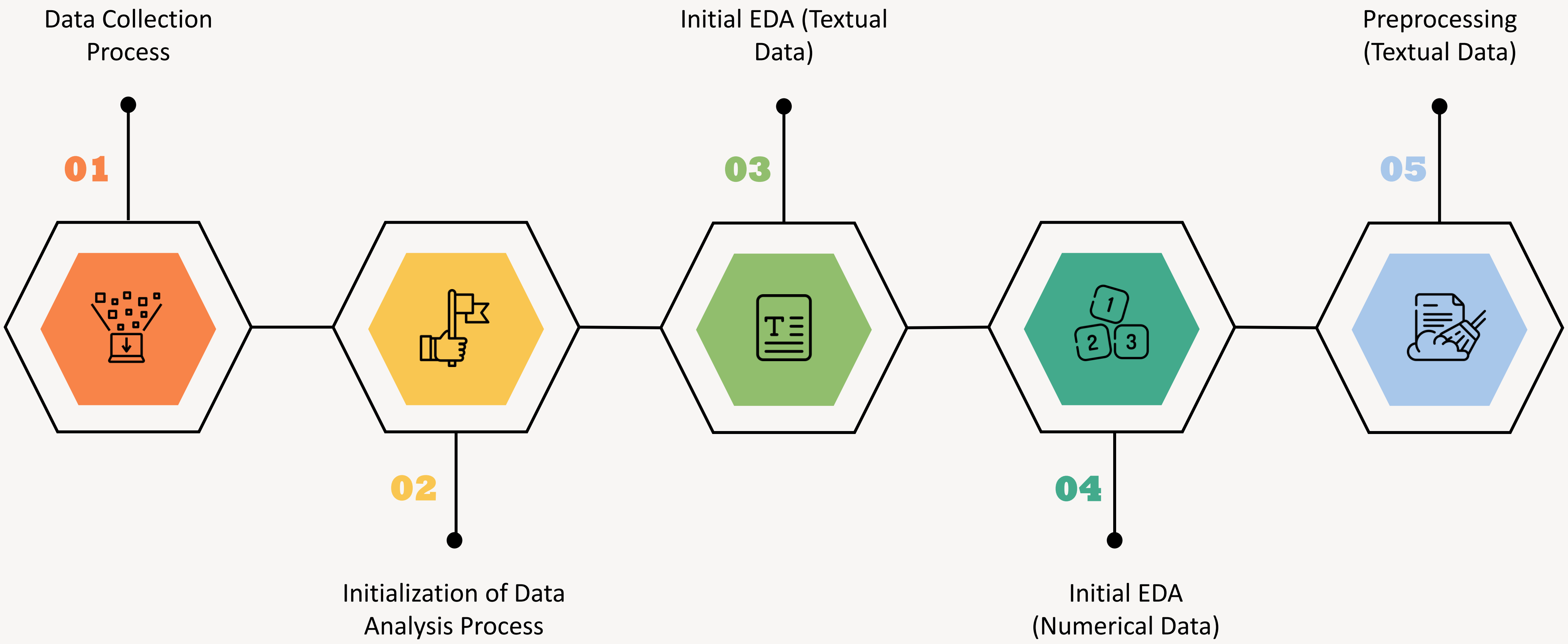
- **Pearson Correlation Between Index and Macroeconomics Indicators: :** This metric is used to evaluate the strength and direction of the linear relationship between the news sentiment index and the five macroeconomics variables (private investment, private consumption, GDP, imports, and exports). A high and positive correlation coefficient indicates that the sentiment index effectively mirrors the fluctuations in macroeconomics variables, thereby showcasing its adequacy as an indicator.

- **Ratio of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE):** This measure assesses the accuracy of the machine learning models used to forecast the five target macroeconomic indicators. Both RMSE and MAE ratios compare the forecasted errors against a baseline model. A ratio lower than 1 signals that the model performs better than the baseline, indicating an improvement in forecast accuracy. Conversely, a ratio higher than 1 would suggest that the model's predictive performance is less accurate than the baseline.

$$Ratio\ RMSE = \frac{RMSE}{RMSE_{AR1}} \qquad Ratio\ MAE = \frac{MAE}{MAE_{AR1}}$$

Note: *AR1* is the baseline model of Ordinary Least Squares with Autoregressive order 1 [OLS-AR(1)].

Implementation



Implementation (cont'd)

06 EDA of Preprocessed Textual Data

07 Computation of News Sentiment Index

08 Nowcasting BCI and CSI

09 Pearson Correlation Between Macroeconomics Variables and News Sentiment Index

10 Modelling Process for Forecasting Activity

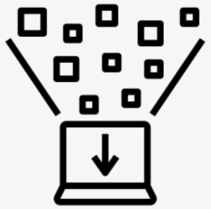
01 Data Collection Process

4.1.1 MIER Dataset

- The Malaysian Institute of Economic Research (MIER) provided survey-based sentiment indices for 2022 and 2023 on a quarterly basis.
- The data was acquired through a purchase and is stored in an Excel file titled "Data BCI_CSI_2022_2023.xlsx".
- Proof of payment for the data purchase is available in *Appendix A*.

4.1.2 Macroeconomics Dataset

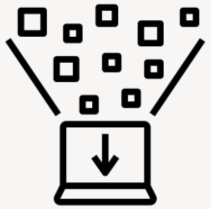
- Macroeconomic data for imports, exports, private investment, private consumption, and GDP for 2022 and 2023 were obtained from OpenDOSM (2024-b) on a quarterly basis.
- Private Final Consumption Expenditure represents private consumption, while Gross Fixed Capital Formation (Private Sector) represents private investment.
- The data reflects the Type of Expenditure at Constant 2015 Prices, expressed as percentage change from the corresponding quarter of the previous year, accounting for seasonality and inflation.
- All values represent Year-over-Year (YoY) growth rates.



01 Data Collection Process (cont'd)

4.1.3 News Articles (Web Scrapping)

- 3,361 articles were successfully scraped from See Hua Daily News, falling short of the initial target of 10,000 articles.
- Several news portals were ruled out due to lack of dedicated finance/business sections (Asia Times, Guangmin Daily, and Overseas Chinese Daily News) or limited accessible historical coverage for 2022-2023 (Kwong Wah Jit Poh, Nanyang Siang Pau, Sin Chew Jit Poh, and China Press).
- Attempts to expand data sources through direct requests or purchases were unsuccessful, despite offering university Research Letters (available in *Appendix B*).
- China Press offered individual article purchases, but costs were prohibitive, estimated at approximately RM30,000 for a two-year dataset.
- ParseHub, a web scraping tool, was used to collect articles from See Hua Daily News, with the process detailed in *Appendix C*.
- The scraped data is stored in a file named "Scraped (See Hua).csv".



02 Initialization of Data Analysis Process

4.2 Initialization Process

- The necessary libraries for data manipulation, statistical modelling, text processing, plotting, and machine learning were imported (available in *Appendix D*).

03 Initial EDA (Textual Dataset)

4.3.1 Loading the Data

- The dataset was initially read from "Scrapped (See Hua).csv" using 'utf-8' encoding after trying various options.

4.3.2 Exploration of Data Structure and Preliminary Cleaning of Dataset

- Initial exploration revealed columns for *article_name*, *article_url*, *article_date*, and *article_article_content*.
- The *article_url* column was dropped, and remaining columns were renamed and reordered for clarity.
- The cleaned DataFrame was saved as "See Hua (New).csv" with 'utf-8' encoding and re-read to ensure correctness.

4.3.3 Exploration of Data Structure of Dataset After Preliminary Cleaning

- The modified dataset contains 3,361 entries across three columns: *Date*, *Title*, and *Content*.
- Summary statistics confirmed no missing entries in any of the columns.

04 Initial EDA (Numeric Dataset)

4.4.1 MIER Dataset

4.4.1.1 Loading the Data

- MIER datasets (BCI and CSI) were loaded from "Data_BCI_CSI_2022_2023.xlsx" using specific sheet names.

4.4.1.2 EDA on the BCI and CSI Data

- Both BCI and CSI datasets contain 8 entries each with no missing values, including columns for Quarter and respective index values.

4.4.1.3 Checking for Outliers

- Outlier analysis using Z-score method showed no outliers in either BCI or CSI datasets.

04 Initial EDA (Numeric Dataset) (cont'd)

4.4.2 Macroeconomics Dataset

4.4.2.1 Loading the Data

- Macroeconomic data was loaded from "Macroeconomics_Data.xlsx", combining imports, exports, GDP, private consumption, and private investment into a single DataFrame.

4.4.2.2 EDA on the Macroeconomics Data

- The combined macroeconomics dataset contains 8 entries for each variable with no missing values.

4.4.2.3 Checking for Outliers

- Outlier analysis for macroeconomic variables also showed no outliers in any of the variables.

05 Preprocessing Steps (Textual Dataset)

4.5.1 Stripping Whitespaces from Column Names

- Column names were stripped of whitespace for standardization.

4.5.2 Converting Chinese Dates to Standard Format

- Chinese dates in the *Date* column were converted to a standard format and set as the DataFrame index.

4.5.3 Text Normalization

- Text in string columns was normalized to handle potential Unicode issues and ensure consistency.

4.5.4 Loading and Displaying Sentiment Dictionary

- The JMZ sentiment dictionary was loaded from "中文金融情感词典_姜富伟等(2021).xlsx", containing positive and negative words, including negators.

4.5.5 Loading Stop Words and Defining Preprocessing Function

- Stop words were loaded from "stopwords-zh.json" to remove non-informative words from the text.
- A preprocessing function was defined to tokenize text and remove stop words.

05 Preprocessing Steps (Textual Dataset) (cont'd)

4.5.6 Concatenating Title and Content, and Performing Textual Preprocessing

- *Title* and *Content* columns were concatenated into a *combined_text* column and tokenized for sentiment analysis and machine learning modelling.

06 EDA of Preprocessed Textual Data

4.6.1 Data Verification

- The first few rows of the preprocessed dataset were displayed to verify the changes made.
- A summary of the preprocessed DataFrame revealed 3,361 entries across four columns: *Title*, *Content*, *combined_text*, and *tokens*.

4.6.2 Saving Preprocessed Textual Data

- The preprocessed data was saved to a CSV file named "preprocessed_data.csv" using 'utf-8' encoding for future use.

07 Computation of News Sentiment Index

4.7.1 Sentiment Analysis

- A sentiment analysis function, *compute_sentiment*, was defined to calculate sentiment scores for each article based on positive and negative word counts.
- The function was applied to the *tokens* column, generating new columns for *sentiment_score*, *positive_count*, *negative_count*, and *total_words*.

4.7.2 Inspecting Sentiment Calculation

- Sample articles were inspected to verify the accuracy of sentiment calculations.

4.7.3 Distribution of Sentiment Scores

- A histogram of sentiment scores showed an approximately normal distribution, indicating balanced sentiment across articles.

4.7.4 Creating the Final DataFrame for Quarterly Sentiment Indices

- Sentiment scores were resampled to compute quarterly averages, aligning with economic indicators.
- A final DataFrame was created with quarterly sentiment indices, with the *Date* column converted to a quarterly format.
- The quarterly sentiment index was saved to "quarterly_sentiment_index.csv" for further analysis.

08 Nowcasting BCI and CSI

4.8.1 Data Preparation for the Nowcasting Activity

- BCI and CSI data from MIER were loaded from "Data BCI_CSI_2022_2023.xlsx" and merged with the quarterly sentiment index.

4.8.2 Plotting Time Series Plot

- A time series plot was created to visually compare the News Sentiment Index with MIER's Business Condition Index and Consumer Sentiment Index.

4.8.3 Regression Analysis for Nowcasting Activity

- Lagged variables for BCI and CSI were created to account for temporal dependencies in the nowcasting activity.
- Regression analysis was performed to evaluate relationships between BCI/CSI values and the quarterly sentiment index using Ordinary Least Squares (OLS) models.

4.8.4 Plotting Actual vs Predicted Values

- Plots comparing actual vs. predicted values for BCI and CSI were generated to visualize nowcasting performance.

08 Nowcasting BCI and CSI (cont'd)

4.8.5 Multicollinearity Check

- Multicollinearity was assessed using the condition number, correlation matrix, and Variance Inflation Factor (VIF) for the BCI nowcasting activity.

09 Pearson Correlation (Target Variables & Index)

4.9.1 Data Preparation for Pearson Correlation Activity

- The quarterly sentiment index data was loaded from "quarterly_sentiment_index.csv" and merged with the macroeconomic data.
- The *Quarter* and *Date* columns were dropped after merging.

4.9.2 Compute Pearson Correlation Coefficients

- Pearson correlation coefficients were computed between the quarterly sentiment index and various macroeconomic variables: imports, exports, GDP, private consumption, and private investment.
- The correlation coefficients were displayed to show the relationships between the sentiment index and these economic indicators.

10 Modelling for Forecasting Activity

4.10.1 Without Hyperparameter Tuning

4.10.1.1 Initialization for the Modelling Process

- A set of functions was defined to calculate Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for model evaluation.
- The modelling process initialized key parameters, including a window size of 4 quarters and forecast horizons of 1, 2, and 3 quarters.
- Macroeconomic variables under study included imports, exports, GDP, private consumption, and private investment.
- Several machine learning models were employed, including OLS Regression, LASSO, Ridge, SVR, Random Forest Regressor, and XGBoost.

4.10.1.2 Rolling Window Approach

- A rolling window approach was implemented to train and test the models, with OLS-AR(1) serving as a benchmark.

4.10.1.3 Display of Results

- Results were visualized through plots of RMSE and MAE ratios for each macroeconomic variable across different models and forecast horizons.

10 Modelling for Forecasting Activity (cont'd)

4.10.2 With Hyperparameter Tuning

4.10.2.1 Initialization for the Modelling Process

- A comprehensive function was developed to compute both RMSE and MAE for model evaluation.
- The modelling process maintained consistent parameters: a 4-quarter window size and forecast horizons of 1, 2, and 3 quarters.
- The study focused on key macroeconomic variables: imports, exports, GDP, private consumption, and private investment.
- Pipelines were utilized to standardize data and apply regression algorithms, ensuring consistent data preprocessing.

4.10.2.2 Hyperparameter Tuning with Grid Search

- Hyperparameter tuning was implemented using grid search with 5-fold cross-validation for LASSO, Ridge, SVR, Random Forest, and XGBoost models.

10 Modelling for Forecasting Activity (cont'd)

4.10.2.3 Use the Best Models for Rolling Window Approach

- A rolling window approach was employed using the best-tuned models, with OLS-AR(1) serving as a benchmark for performance comparison.

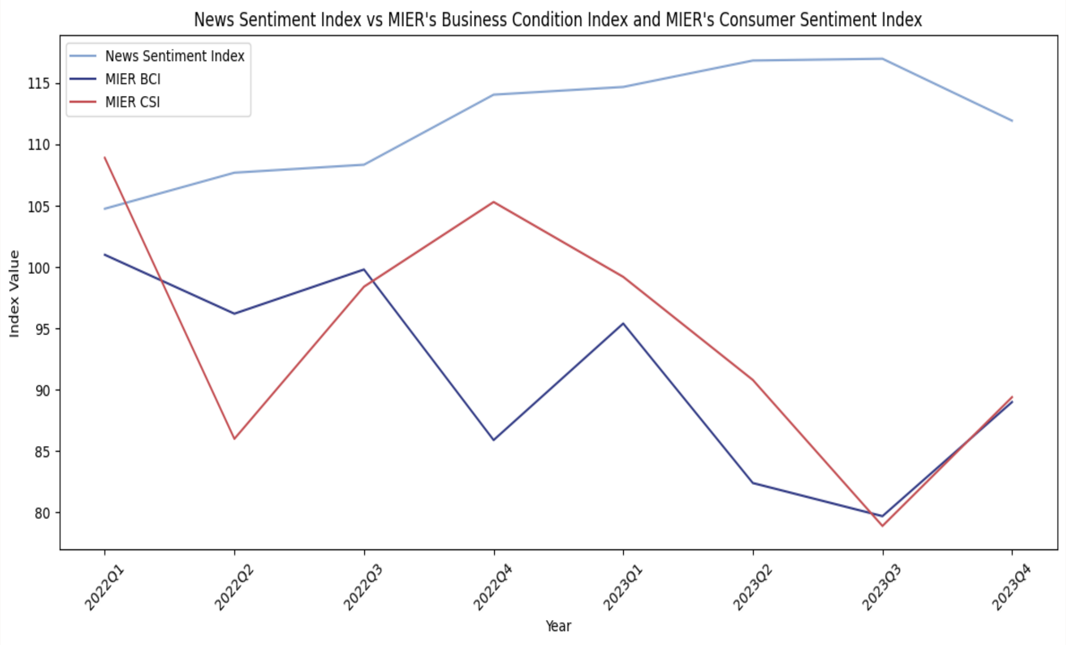
4.10.2.4 Display of Results

- Results were visualized through plots of RMSE and MAE ratios, and average performance metrics were calculated for each model, variable, and horizon.
- The best parameters identified during tuning were recorded and displayed for reference and reproducibility.

Results and Analysis

5.1 Nowcasting the BCI and CSI using the News Sentiment Index

5.1.1 Time Series Plot

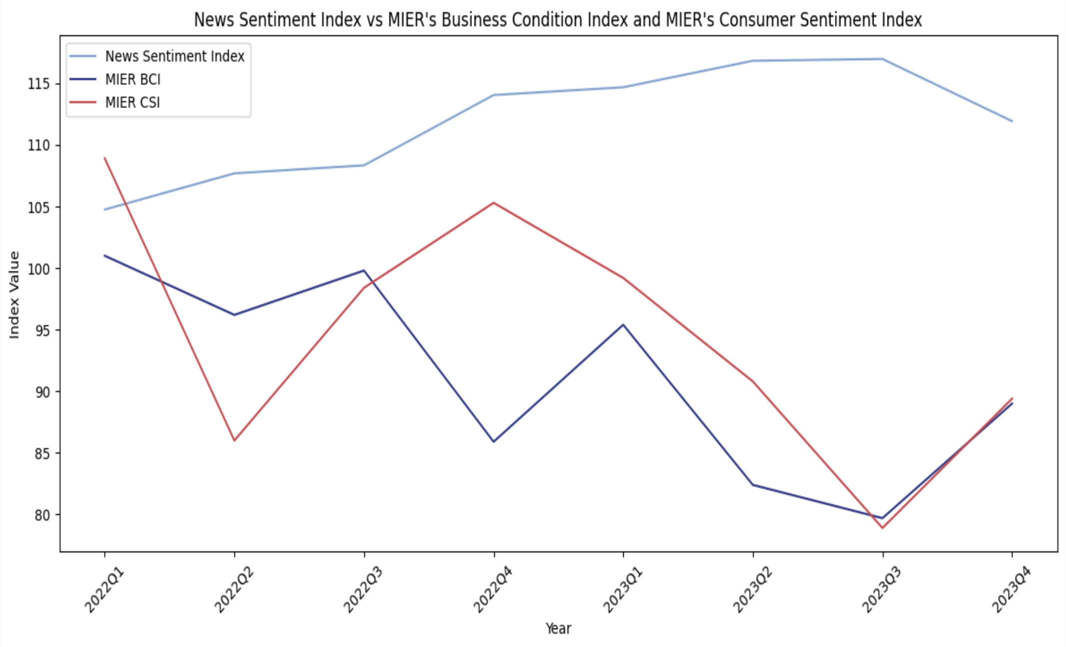


- The News Sentiment Index demonstrates an overall upward trend from 2022 Q1 to 2023 Q4, indicating a general increase in positive sentiment.
- MIER's BCI and CSI show more volatility compared to the News Sentiment Index.
- The News Sentiment Index and BCI exhibit some alignment from 2022 Q1 to 2022 Q3, both declining and then recovering.
- Post-2022 Q3, the BCI sharply declines until 2023 Q3, diverging from the News Sentiment Index, suggesting a moderately weak correlation.

Results and Analysis

5.1 Nowcasting the BCI and CSI using the News Sentiment Index

5.1.1 Time Series Plot (cont'd)



- The CSI initially starts higher than the News Sentiment Index but experiences steep drops and partial recoveries, indicating a very weak correlation with the News Sentiment Index.
- **Overall, the analysis suggests a moderately weak correlation between the News Sentiment Index and BCI, and a very weak correlation with the CSI.**

Results and Analysis

5.1.2 Regression Output for Nowcasting MIER's BCI

OLS Regression Results

Dep. Variable:	BCI Values	R-squared:	0.697
Model:	OLS	Adj. R-squared:	0.545
Method:	Least Squares	F-statistic:	4.592
Date:	Fri, 26 Jul 2024	Prob (F-statistic):	0.0921
Time:	14:54:19	Log-Likelihood:	-19.399
No. Observations:	7	AIC:	44.80
Df Residuals:	4	BIC:	44.64
Df Model:	2		
Covariance Type:	nonrobust		

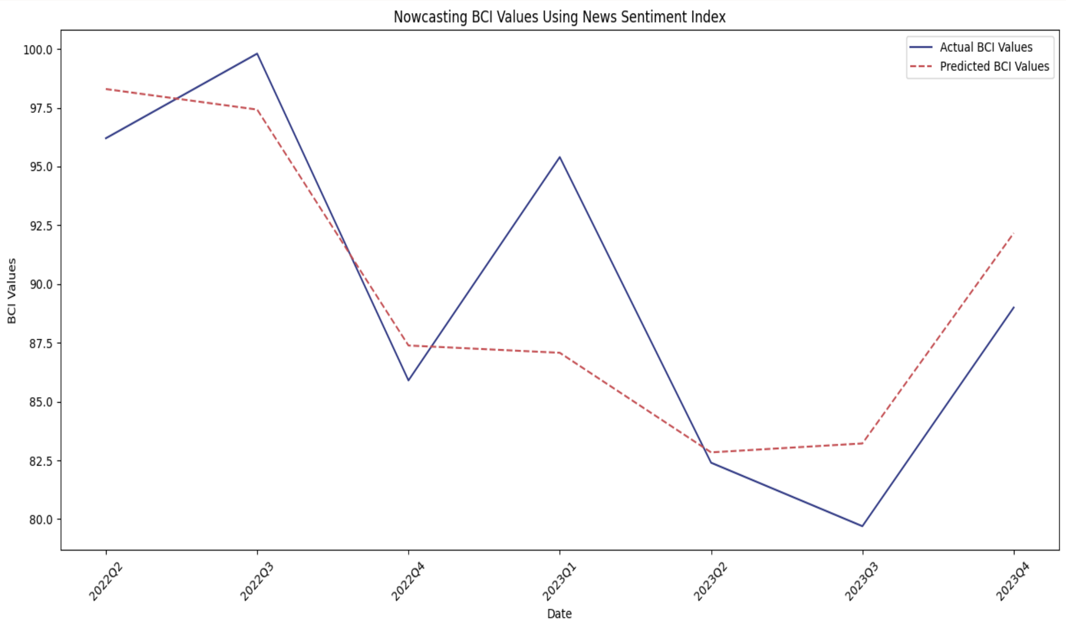
	coef	std err	t	P>|t|	[0.025	0.975]
const	289.5625	81.926	3.534	0.024	62.099	517.026
BCI_Lag	-0.0541	0.265	-0.204	0.848	-0.791	0.682
quarterly_sentiment_index	-1.7254	0.609	-2.834	0.047	-3.416	-0.035

Omnibus:	nan	Durbin-Watson:	2.089
Prob(Omnibus):	nan	Jarque-Bera (JB):	1.849
Skew:	1.250	Prob(JB):	0.397
Kurtosis:	3.303	Cond. No.	6.16e+03

- The regression model shows a moderately strong fit with an R-squared of 0.697, explaining about 69.7% of BCI variance.
- The model is marginally significant at the 10% level (F-statistic: 4.592, p-value: 0.0921).
- The lagged BCI variable is not statistically significant, suggesting little impact on current BCI values.
- The quarterly sentiment index variable is statistically significant at the 5% level, but unexpectedly shows a negative relationship with BCI.
- No significant autocorrelation is present in the residuals (Durbin-Watson statistic: 2.089).
- The high condition number (6.16e+03) raises concerns about potential multicollinearity or numerical issues in the model.

Results and Analysis

5.1.2 Regression Output for Nowcasting MIER's BCI (cont'd)



- The nowcasting plot shows that predicted BCI values follow the general trend of actual values but with less volatility and potential overestimation.
- **Overall, the model demonstrates some explanatory power but also highlights significant issues, including the unexpected negative relationship between sentiment index and BCI, and the high condition number.**

Results and Analysis

5.1.3 Regression Output for Nowcasting MIER's CSI

OLS Regression Results

Dep. Variable:	CSI Values	R-squared:	0.031
Model:	OLS	Adj. R-squared:	-0.454
Method:	Least Squares	F-statistic:	0.06367
Date:	Fri, 26 Jul 2024	Prob (F-statistic):	0.939
Time:	14:54:25	Log-Likelihood:	-24.644
No. Observations:	7	AIC:	55.29
Df Residuals:	4	BIC:	55.13
Df Model:	2		
Covariance Type:	nonrobust		

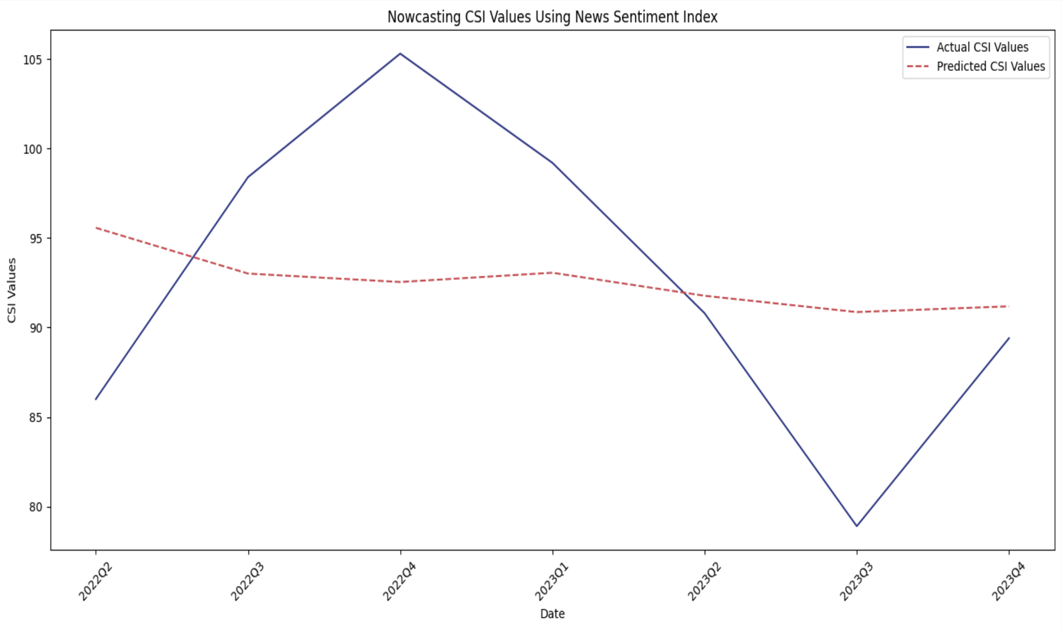
	coef	std err	t	P>|t|	[0.025	0.975]
const	117.3405	137.175	0.855	0.441	-263.517	498.198
CSI_Lag	0.1029	0.413	0.249	0.816	-1.045	1.251
quarterly_sentiment_index	-0.3063	1.171	-0.261	0.807	-3.558	2.946

Omnibus:	nan	Durbin-Watson:	1.274
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.394
Skew:	-0.034	Prob(JB):	0.821
Kurtosis:	1.840	Cond. No.	4.96e+03

- The regression model for the Consumer Sentiment Index (CSI) demonstrates a weak fit with an R-squared of only 0.031, explaining just 3.1% of CSI variance.
- The model is not statistically significant (F-statistic: 0.06367, p-value: 0.939), indicating that the predictors do not effectively explain CSI variability.
- Neither the lagged CSI variable nor the quarterly sentiment index variable are statistically significant in predicting CSI values.
- A slight positive autocorrelation is present in the residuals (Durbin-Watson statistic: 1.274).
- The high condition number (4.96e+03) suggests potential multicollinearity or numerical issues in the model.

Results and Analysis

5.1.3 Regression Output for Nowcasting MIER's CSI (cont'd)



- The nowcasting plot reveals that predicted CSI values follow a general downward trend, failing to capture the significant fluctuations in actual CSI values.
- **Overall, the model demonstrates very little explanatory power and fails to accurately predict CSI values, rendering further checks for multicollinearity or numerical issues unnecessary.**

Results and Analysis

5.2 Forecasting the BCI and CSI using the News Sentiment Index

5.2.1 Pearson Correlation (Macroeconomic Variables and Index)

```
Pearson Correlation Coefficients with Quarterly Sentiment Index:
quarterly_sentiment_index    1.000000
Imports                     -0.889319
Exports                     -0.817534
GDP                         -0.494618
Private Consumption         -0.532231
Private Investment           0.052724
Name: quarterly_sentiment_index, dtype: float64
```

- The Pearson correlation analysis reveals strong negative correlations between the quarterly sentiment index and imports (-0.889319) and exports (-0.817534).
- GDP (-0.494618) and private consumption (-0.532231) show moderate negative correlations with the sentiment index.
- Private investment exhibits a very weak positive correlation (0.052724) with the quarterly sentiment index.
- These results suggest that as the sentiment index increases, imports, exports, GDP, and private consumption tend to decrease, while private investment remains largely unaffected.

Results and Analysis

5.2.1 Pearson Correlation (Macroeconomic Variables and Index) (cont'd)

```
Pearson Correlation Coefficients with Quarterly Sentiment Index:
quarterly_sentiment_index    1.000000
Imports                     -0.889319
Exports                     -0.817534
GDP                         -0.494618
Private Consumption         -0.532231
Private Investment           0.052724
Name: quarterly_sentiment_index, dtype: float64
```

- The observed negative correlations are unexpected, as a higher sentiment index typically suggests more positive news sentiment, which would generally be expected to correlate positively with economic activity indicators.
- The weak positive correlation with private investment is also contrary to expectations, as more positive sentiment would typically be associated with higher investment levels.
- **Overall, the News Sentiment Index shows unexpected correlations with economic indicators, revealing complex relationships between news sentiment and Malaysia's economy.**

Results and Analysis

5.2.2 Performance of Machine Learning Model Without Hyperparameter Tuning

FINDINGS:

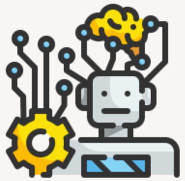
- LASSO consistently performed well for GDP, with ratios below 1 across all three forecasting horizons.
- For private consumption, LASSO, Random Forest, and XGBoost models all met the criterion of having ratios below 1 for all horizons.
- Private investment showed consistent performance with both LASSO and Ridge models.
- No models achieved RMSE and MAE ratios below 1 across all horizons for imports and exports variables.
- The criterion for selecting robust models was based on having at least three best performances out of the five macroeconomic variables.
- LASSO emerged as the standout model, excelling in predicting GDP, private consumption, and private investment for all forecasting horizons.
- This selection method prioritizes models demonstrating broader and more consistent predictive capability across most analysed variables.

Results and Analysis

5.2.3 Performance of Machine Learning Model With Hyperparameter Tuning

FINDINGS:

- For imports, exports, and GDP variables, no models met the criterion of having all ratios below 1.
- The Random Forest model showed strong predictive performance for private consumption, with all ratios below 1 across all horizons.
- For private investment, LASSO and SVR models demonstrated consistent performance with ratios below 1 across all horizons.
- No model achieved at least three best performances across the five variables studied.



Discussions

6.1 Nowcasting Findings

- The Consumer Sentiment Index (CSI) nowcasting model showed poor performance, with a low R-squared value and non-significant F-statistic, indicating weak predictive power.
- Multicollinearity testing for the Business Condition Index (BCI) regression showed no issues, with VIF values below the threshold of 10.
- The small sample size (7 observations over 2 years) may have led to numerical instability in the regression analysis.
- The quarterly news sentiment index was found to be significant in predicting MIER's Business Condition Index (BCI), consistent with previous research.
- An unexpected negative coefficient for the quarterly sentiment index was observed, possibly due to other influencing factors or the small sample size.
- Significant events during 2022-2023, such as the ongoing Russian-Ukraine war, Malaysia's hung parliament situation, and the formation of a new coalition government, may have influenced the BCI despite positive news sentiment.
- These political and global events likely created uncertainty, affecting business confidence adversely and potentially explaining the negative relationship between news sentiment and BCI.
- Further research with a larger sample size and additional variables, including political stability indicators and global economic conditions, is recommended to better understand the relationship between news sentiment and business conditions.

Discussions

6.1 Nowcasting Findings (cont'd)

- The findings highlight the potential of news sentiment as a predictive tool for business sentiment, while also emphasizing the need for more comprehensive modelling that accounts for major political and global events.

Discussions

6.2 Forecasting Findings

- Unexpected negative correlations were found between the quarterly news sentiment index and several macroeconomic variables (imports, exports, GDP, and private consumption).
- Private investment showed a weak positive correlation with the quarterly sentiment index, aligning more closely with expectations.
- These findings differ from Chong et al. (2021), who found generally positive correlations between news sentiment and economic indicators.
- Machine learning models without hyperparameter tuning performed better than tuned models in the forecasting activity.
- The LASSO model emerged as the most robust, effectively predicting GDP, private investment, and private consumption across all forecasting horizons.
- Chong et al. (2021) found news sentiment consistently forecasted private investment growth better than a benchmark model, particularly for 2-3 quarters ahead.
- The current study's findings partially align with Chong et al. (2021) regarding private investment prediction but differ in identifying LASSO as robust for other variables.
- The analysis highlights the complexities of using news sentiment indices to forecast macroeconomic variables and suggests areas for future research.

Conclusions

Objectives

VS

Findings

01

To evaluate the ability of Mandarin-based news sentiment indices in predicting sentiments reflected by the BCI and CSI figures.

- The Consumer Sentiment Index (CSI) nowcasting model showed poor performance, aligning with previous research on the weak relationship between news sentiment and consumer sentiment.
- The quarterly news sentiment index was significant in predicting the Business Condition Index (BCI), reinforcing its potential as a forecasting tool for business sentiment.

Conclusions

Objectives	Findings
01 To evaluate the ability of Mandarin-based news sentiment indices in predicting sentiments reflected by the BCI and CSI figures.	▪ The Consumer Sentiment Index (CSI) nowcasting model showed poor performance, aligning with previous research on the weak relationship between news sentiment and consumer sentiment. ▪ The quarterly news sentiment index was significant in predicting the Business Condition Index (BCI), reinforcing its potential as a forecasting tool for business sentiment.

Conclusions

Objectives	Findings
02 To compare the performance of the Mandarin text-derived news sentiments to that of the English text-derived news sentiments reported in previous published work.	▪ The LASSO model emerged as the most robust, effectively predicting GDP, private investment, and private consumption across all forecasting horizons. ▪ The study partially aligns with previous research, particularly in the predictive power of news sentiment for private investment.

Conclusions

Objectives

VS

Findings

02

To compare the performance of the Mandarin text-derived news sentiments to that of the English text-derived news sentiments reported in previous published work.

- The LASSO model emerged as the most robust, effectively predicting GDP, private investment, and private consumption across all forecasting horizons.
- The study partially aligns with previous research, particularly in the predictive power of news sentiment for private investment.

Conclusions

Objectives	Findings
03 To evaluate the correlation between the four demand-side components of GDP and the news sentiment index.	▪ Strong negative correlations were found between the sentiment index and imports (-0.889) and exports (-0.817), while moderate negative correlations were observed with GDP (-0.495) and private consumption (-0.532). Private investment showed a very weak positive correlation (0.053) with the sentiment index.

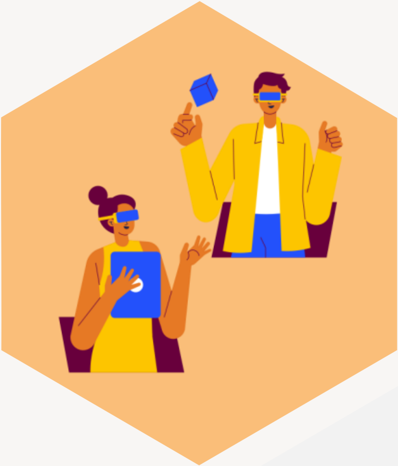
VS

Conclusions

Objectives	Findings
03 To evaluate the correlation between the four demand-side components of GDP and the news sentiment index.	▪ Strong negative correlations were found between the sentiment index and imports (-0.889) and exports (-0.817), while moderate negative correlations were observed with GDP (-0.495) and private consumption (-0.532). Private investment showed a very weak positive correlation (0.053) with the sentiment index.

Limitations

- Reliance on a single news source (See Hua Newspaper) limited the capture of diverse economic sentiments.
- The short study period (2022-2023) may have affected the robustness of the findings by not accounting for longer-term economic cycles or sentiment trends.



Future Work

- Incorporate a broader range of news sources to capture more diverse economic sentiments.
- Extend the data period to include more years, providing a more comprehensive view of long-term economic cycles and trends.
- Include additional influential variables such as political stability, global economic conditions, and domestic policy changes to enhance the predictive models' robustness and accuracy.

In conclusion

Mandarin news sentiment demonstrated potential as a valuable tool in economic forecasting for Malaysia, despite the shorter study period compared to previous English-based research.

References

- Abir Masmoudi, Hamdi Jamila, & Lamia Hadrich Belguith. (2021). Deep learning for sentiment analysis of Tunisian dialect. *Computación y Sistemas, 25*(1), 129–148. doi: 10.13053/CyS-25-1-3472
- Agu, S. C., Onu, F. U., Ezemagu, U. K., & Oden, D. (2022). Predicting gross domestic product to macroeconomic indicators. *Intelligent Systems with Applications*, *14*, 200082. https://doi.org/10.1016/j.iswa.2022.200082
- Al Shamsi, A. A., & Sherief Abdallah. (2022). Sentiment Analysis of Emirati Dialect. *Big Data and Cognitive Computing*, *6*(2), 57–57. https://doi.org/10.3390/bdcc6020057
- Ashwin, J., Eleni, K., Lorena, S. (2021, November). *Nowcasting euro area GDP with news sentiment: A tale of two crises.* (ECB Working Paper, No. 2616). doi:10.2866/240669
- Barbaglia, L., Consoli, S., & Sebastiano Manzan. (2024). Forecasting GDP in Europe with textual data. *Journal of Applied Econometrics, 39*(2), 338-355. https://doi.org/10.1002/jae.3027
- Barbaglia, L., Frattarolo, L., Luca Onorante, Filippo Maria Pericoli, Ratto, M., & Luca Tiozzo Pezzoli. (2023). Testing big data in a big crisis: Nowcasting under Covid-19. *International Journal of Forecasting*, *39*(4), 1548–1563. https://doi.org/10.1016/j.ijforecast.2022.10.005
- Chong, E., Ho, C. C., Ong, Z. F., & Ong, H. H. (2021). *Using news sentiment for economic forecasting: A Malaysian case study*. https://www.bis.org/ifc/publ/ifcb57_17.pdf
- Chu, B., & Qureshi, S. (2023). Comparing out-of-sample performance of machine learning methods to forecast U.S. GDP growth. *Computational Economics*, *62*, 1567-1609. https://doi.org/10.1007/s10614-022-10312-z

References (cont'd)

- Correa, R., Garud, K., Londono-Yarce, J.-M., & Mislang, N. (2017). Constructing a dictionary for financial stability. IFDP Notes, 2017(33), 1–7. https://doi.org/10.17016/2573-2129.33
- Diaz, G., & fseasy. (2020). *stopwords-iso / stopword-zh*. GitHub. https://github.com/stopwords-iso/stopwords-zh
- Dauphin, J-F., Dybczak, K., Maneely, Sanjani, M. T., Suphaphiphat, N., Wang, Y., & Zhang, H. (2022, March). *Nowcasting GDP-A scalable approach using DFM, machine learning and novel data, applied to European economies.* (Working Paper No. 2022/052). https://www.imf.org/en/Publications/WP/Issues/2022/03/11/Nowcasting-GDP-A-Scalable-Approach-Using-DFM-Machine-Learning-and-Novel-Data-Applied-to-513703
- Gupta, V., Jain, N., Shubhm, S., Madan, A., Chaudhary, A., & Qin, X. (2021). Toward integrated CNN-based sentiment analysis of tweets for scarce-resource language—Hindi. *ACM Transactions on Asian Low-Resource Language Information Processing, 20*(5), 80(1)-80(23). https://doi.org/10.1145/3450447
- Hegde, A., Chakravarthi, B. R., Shashirekha1, H. L., Ponnusamy, R., Navaneethakrishnan, S. C., Kumar, L. S., Thenmozhi, D., Karunakar, M., Sriram, S., & Aymen. S. (2023). Findings of the shared task on sentiment analysis in Tamil and Tulu. *Third Workshop on Speech and Language Technologies for Dravidian Languages,* 64-71. https://aclanthology.org/2023.dravidianlangtech-1.pdf
- Huang, A., Wu, W., & Yu, T. (2020). Textual analysis for China's financial markets: a review and discussion. *China Finance Review International, 10*(1), 1-15. https://doi.org/10.1108/CFRI-08-2019-0134
- Jasni, S. K., Aris, F. E. M., & Jamilat, V. S. (2022). *Nowcasting Malaysia's GDP with machine learning*. Department of Statistics Malaysia. https://drive.google.com/file/d/1o0InpFfcqyYlBbk4YaD9t0ntvWCKxff5/view

References (cont'd)

- Jiang, F., Lee, J., Martin, X., & Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics*, *132*(1), 126–149. https://doi.org/10.1016/j.jfineco.2018.10.001

- Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2022). Making text count: Economic forecasting using newspaper text. *Journal of Applied Econometrics*, *37*(5), 896–919. https://doi.org/10.1002/jae.2907

- Kant, D., Pick, A., & Winter, J. D. (2022, November). *Nowcasting GDP using machine learning algorithms.* (DNB Working Paper No. 754). https://www.dnb.nl/media/kq4pe4cr/working_paper_no_754.pdf

- Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Mahalingam, P., Subramanian, M., Prasanna Kumar Kumaresan, & Ruba Priyadharshini. (2022). An analysis of machine learning models for sentiment analysis of Tamil code-mixed data. *Computer Speech & Language*, *76*, 101407–101407. https://doi.org/10.1016/j.csl.2022.101407

- Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. The Journal of Finance, 66(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

- Martin, L. C. (2019, August). *Machine learning vs traditional forecasting methods: An application to South African GDP*. (Stellenbosch Economic Working Papers WP12/2019). https://ideas.repec.org/p/sza/wpaper/wpapers326.html

- Moudjari, L., & Akli-Astouati, K. (2020). An Experimental Study on Sentiment Classification of Algerian Dialect Texts. Procedia Computer Science, 176, 1151–1159. https://doi.org/10.1016/j.procs.2020.09.111

References (cont'd)

- Muchisha, N. D., Tamara, N., Andriansyah, & Soleh, A. M. (2021). Nowcasting Indonesia's GDP using machine learning algorithms. *Indonesian Journal of Statistics and Its Applications*, *5*(2), 355-368. https://mpra.ub.uni-muenchen.de/105235/1/MPRA_paper_105235.pdf

- Muhammad Fakhrur Razi Abu Bakar, Norisma Idris, Liyana Shuib, & Norazlina Khamis. (2020). Sentiment analysis of noisy Malay text: State of art, challenges and future work. *IEEE Access*, *8*, 24687–24696. https://doi.org/10.1109/access.2020.2968955

- Nakazawa, T. (2022, July). *Constructing GDP nowcasting models using alternative data*. (Bank of Japan Working Paper Series No.22-E-5). https://www.boj.or.jp/en/research/wps_rev/wps_2022/data/wp22e09.pdf

- Nothman, J., Qin, H., & Yurchak, R. (2019). *Stop word lists in free open-source software packages*. https://doi.org/10.18653/v1/w18-2502

- Ong, J. Y., Muhammad Mun'im Zabidi, Norhafizah Ramli, & Sheikh, U. U. (2020). Sentiment analysis of informal Malay tweets with deep learning. *IAES International Journal of Artificial Intelligence (IJ-AI), 9*(2), 212-220. DOI:10.11591/ijai.v9.i2.pp212-220

- OpenDOSM. (2024-b). *[4Q 2023] Gross domestic product*. https://open.dosm.gov.my/publications/gdp_2023-q4

- Rana, T. A., Shahzadi, Rana, T., K., Arshad, A., & Tubishat, M. (2021). An unsupervised approach for sentiment analysis on social media short text classification in Roman Urdu. *ACM Transactions on Asian Low-Resource Language Information Processing, 21*(2), 28(1)-28(16). https://doi.org/10.1145/3474119

References (cont'd)

- Richardson, A., Mulder, T. V. F., & Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, *37*(2), 941-948. https://doi.org/10.1016/j.ijforecast.2020.10.005
- Shapiro, A. H., Sudhof, M., & Wilson, D. (2022). Measuring news sentiment. *Journal of Econometrics,* 228(2022), 221-243. https://linkinghub.elsevier.com/retrieve/pii/S0304407620303535
- Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications, 34(4), 2622–2629. https://doi.org/10.1016/j.eswa.2007.05.028
- Thorsrud, L. A. (2016, December). Words are the new numbers: A newsy coincident index of business cycles. (Working Paper, No. 21/2016). https://hdl.handle.net/11250/2495606
- Wang, D., & Alfred, R. (2020). A review on sentiment analysis model for Chinese Weibo text, 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), 456-463, doi: 10.1109/AEMCSE50948.2020.00105.
- Wei, Z., Liu, W., Zhu, G., Zhang, S., & Hsieh, M-Y. (2022) Sentiment classification of Chinese Weibo based on extended sentiment dictionary and organisational structure of comments. Connection Science, 34(1), 409-428. DOI: 10.1080/09540091.2021.2006146
- Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment analysis for e-commerce product reviews in Chinese based on sentiment lexicon and deep learning. IEEE Access, 8, 23522–23530. https://doi.org/10.1109/access.2020.2969854
- Zhang, Q., He, N., & Xu, H. (2023). Nowcasting Chinese GDP in a data-rich environment: Lessons from machine learning algorithms. *Economic Modelling*, *122*, 106204. https://doi.org/10.1016/j.econmod.2023.106204

References (cont'd)

- Zhang, S., Wei, Z., Wang, Y., & Liao, T. (2018). Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary. Future Generation Computer Systems, 81, 395–403. https://doi.org/10.1016/j.future.2017.09.048'=