1. Synthetic Dataset Creation

Objective:

The first major task was to create a synthetic dataset that mimics the structure and scale of real-world financial news data. This dataset would serve as the input for constructing sentiment indices, a key element in the paper's methodology.

Process:

1. **Dataset Structure:**
    - A dataset with **1 million data points** was generated to simulate financial news articles. Each row represents a "news article," with columns for the date of publication, the news source, key economic keywords, and sentiment scores.
    - Attributes of the dataset include:
        - **Dates:** Covering a wide range of years to reflect realistic economic cycles.
        - **Sources:** Simulated from prominent financial news outlets such as *Bloomberg*, *Reuters*, *The Economist*, and *The Wall Street Journal*.
        - **Keywords:** Included terms like "inflation," "GDP," "employment," and "recession" to capture macroeconomic themes.
2. **Simulating Sentiment Scores:**
    - Sentiment scores were synthetically generated using random sampling methods, informed by plausible distributions of positive and negative sentiment.
    - The scores range from highly pessimistic (negative) to highly optimistic (positive), reflecting the range of market sentiment that might be observed in real data.

Date	Source	Headline	Content
2003-12-13	The Wall Street Journal	Employment rates declines	Analysts predict that GDP will affect upcoming fiscal policies. The GDP shows a signif
2016-07-03	Reuters	Federal Reserve policies worsens	The relationship between fiscal policy and monetary policy is widely debated. Historica
2002-06-19	The Wall Street Journal	Employment rates surges	Analysts predict that economic growth will affect upcoming fiscal policies. Recent data
2008-02-04	The Wall Street Journal	Federal Reserve policies surges	The impact of GDP on global trade remains a major concern. Historical trends in GDP
2008-03-17	The Wall Street Journal	Recession fears improves	Uncertainty around fiscal policy creates challenges for policymakers. Recent data on f
2002-10-13	The Wall Street Journal	GDP growth declines	The recession shows a significant change due to market conditions. Stakeholders are
2005-02-11	Financial Times	GDP growth remains steady	Stakeholders are closely monitoring developments in recession. Stakeholders are clos
2011-08-18	The Wall Street Journal	Employment rates remains steady	Uncertainty around recession creates challenges for policymakers. Government policie
2018-01-24	The Economist	Economic recovery worsens	Government policies addressing economic growth are under scrutiny. The relationship
2006-08-28	The Wall Street Journal	Economic recovery improves	Historical trends in economic growth reveal cyclical patterns of economic activity. Gov

Outcome:

- The resulting dataset provides a robust foundation for constructing sentiment indices using methodologies described in the paper.

- The large size (1 million data points) ensures statistical reliability when calculating sentiment indices.

	Keywords
dict that GDP will affect up	inflation
emains a major concern. I	fiscal policy
coming fiscal policies. Gc	recession
erts believe GDP may influ	inflation
nomic activity. Recent data	GDP
y debated. Analysts predi	employment
ysts predict that recessior	recession
ebated. The relationship b	economic growth
ade remains a major conc	GDP
challenges for policymake	fiscal policy

2. Acquiring GDP Data

The next step was to gather real-world macroeconomic data to serve as the dependent variable in our econometric models. Specifically, we focused on quarterly real GDP data from multiple vintages.

observation_date	GDPC1_20241030	GDPC1_20241127
2002-01-01	14372.785	14372.785
2002-04-01	14460.848	14460.848
2002-07-01	14519.633	14519.633
2002-10-01	14537.580	14537.580
2003-01-01	14614.141	14614.141
2003-04-01	14743.567	14743.567
2003-07-01	14988.782	14988.782

Process:

1. **Data Source:**
    - The ALFRED database (Archival Federal Reserve Economic Data) was chosen as the primary source for GDP data.

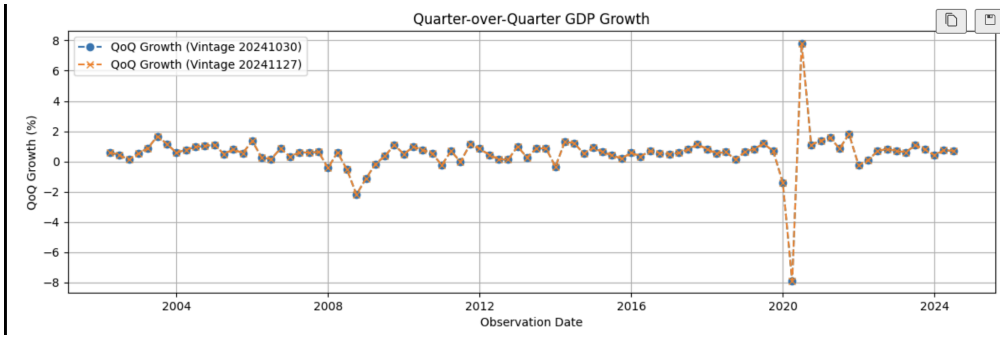
    - ALFRED provides access to vintage data, allowing us to observe how GDP estimates have been revised over time.
2. **Data Retrieval:**
    - Quarterly real GDP values were retrieved for two specific vintages: **October 2024** and **November 2024.**
    - The data was aligned by observation date and reference quarter to facilitate comparisons across vintages.
3. **GDP Growth Rate Calculation:**
    - This transformed raw GDP levels into a more interpretable metric for economic performance.

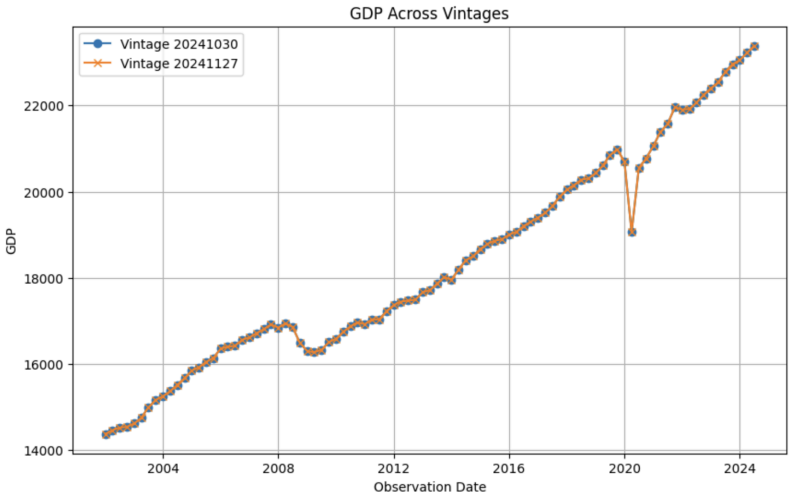
Insights Gained:

- Examining GDP across vintages highlighted the significance of real-time data revisions.
- These revisions emphasized the importance of incorporating lagged predictors into econometric models, as real-time estimates are inherently subject to updates.

3. Exploratory Data Analysis (EDA)

To explore and understand the structure of the GDP data and its relationship to sentiment indices, paving the way for econometric modeling.





Process:

1. **EDA on GDP Data:**
    - Quarterly GDP values were plotted to observe trends and seasonal patterns.
    - Growth rates were analyzed to detect periods of economic expansion and contraction.
2. **Cross-Vintage Analysis:**
    - Differences between GDP vintages were analyzed to understand the magnitude and direction of revisions.
    - This analysis confirmed the necessity of using multiple vintages to model real-time decision-making processes.
3. **Integration with Sentiment Indices:**
    - Sentiment indices were merged with GDP data based on observation dates and quarters.
    - The merged dataset was inspected for missing values, inconsistencies, and alignment issues.
    -

4. Sentiment Analysis with FIGAS

To replicate the construction of sentiment indices using the Financial News Aggregated Sentiment (FIGAS) approach, as described in the paper.

Date	Source	Keywords	Aggregate_Sentiment
2021-09-09	Bloomberg	inflation	-2.0
2013-08-29	The Economist	recession	0.0
2008-04-21	Reuters	GDP	0.0
2015-08-02	The Wall Street Journal	inflation	-1.5
2017-10-10	The Wall Street Journal	economic growth	-0.5
2017-12-30	Financial Times	employment	-0.42857142857142900
2006-10-01	The Economist	recession	0.0
2016-05-29	Bloomberg	fiscal policy	0.0
2014-03-26	Reuters	employment	-1.8
2010-05-03	The Wall Street Journal	fiscal policy	0.0
2015-07-19	Financial Times	fiscal policy	0.0
2008-08-20	The Economist	economic growth	-1.2857142857142900
2005-02-27	Bloomberg	economic growth	-0.6
2007-10-16	The Economist	employment	-1.7142857142857100

Figas sentiment results

Process:

1. **Execution:**

- FIGAS was applied to the synthetic dataset to compute sentiment scores for each keyword, source, and time period.
- Aggregated sentiment scores were calculated by averaging daily scores within each quarter.

Quarter	Sentiment_Index	Normalized_Sentiment_Index
2002Q1	-0.5682666823955020	0.0217015841111413
2002Q2	-0.5496655143792440	0.08693002362543270
2002Q3	-0.5265162836218830	0.16810708718074900
2002Q4	-0.5744553144596160	0.0
2003Q1	-0.5307268381391190	0.15334199821380200
2003Q2	-0.5274860394146680	0.16470645895041100

Sentiment index by quarter

2. **Technical Challenges:**
   - The computation process was resource-intensive, given the size of the dataset (1 million rows).
   - Running the FIGAS script took **over 24 hours** on the lab's server to complete due to the complexity of aggregating and normalizing sentiment scores for a million articles.
3. **Output:**
   - Quarterly sentiment indices were produced, with normalization applied to ensure comparability over time.

5. Integration of Sentiment Indices with GDP Data

To merge sentiment indices with GDP data, creating a comprehensive dataset for econometric modeling.

Process:

1. **Data Merging:**
   - Sentiment indices were aligned with GDP data based on quarters.
   - Additional variables, such as lagged sentiment indices and normalized indices, were computed and included.
2. **Outcome:**
   - A finalized dataset was created, containing:

        - Quarterly GDP levels and growth rates.
        - Sentiment indices.
        - Lagged sentiment indices.
    - This dataset serves as the input for econometric modeling.

	GDPC1_20241030	GDPC1_20241127	Quarter	Sentiment_Index	Normalized_Sentiment_Index	GDP_Growth_Rate	Lagged_Sentiment_Index
2	14460.848	14460.848	2002Q2	-0.5496655143792440	0.0869300236254326	0.6127065840058070	0.0217015841111413
3	14519.633	14519.633	2002Q3	-0.5265162836218830	0.1681070871807490	0.4065114300350950	0.0869300236254326
4	14537.58	14537.58	2002Q4	-0.5744553144596160	0.0	0.12360505255195600	0.1681070871807490
5	14614.141	14614.141	2003Q1	-0.5307268381391190	0.1533419982138020	0.5266419858050540	0.0
6	14743.567	14743.567	2003Q2	-0.5274860394146680	0.1647064589504110	0.885621672871495	0.1533419982138020
7	14988.782	14988.782	2003Q3	-0.5495095145399650	0.0874770659440506	1.663199956971060	0.1647064589504110
8	15162.76	15162.76	2003Q4	-0.5434125993054570	0.1088570280115030	1.160721398176310	0.0874770659440506
9	15248.68	15248.68	2004Q1	-0.5451770425341850	0.1026696811566740	0.5666514539569350	0.1088570280115030
10	15366.85	15366.85	2004Q2	-0.5452166728940920	0.1025307099624	0.7749523237421170	0.1026696811566740
11	15512.619	15512.619	2004Q3	-0.5607795039479420	0.0479567615318529	0.9485938887930920	0.1025307099624
12	15670.88	15670.88	2004Q4	-0.5366098863219320	0.1327120005589860	1.0202081286209600	0.0479567615318529
13	15844.727	15844.727	2005Q1	-0.5210991182141710	0.1871033806300080	1.1093633541958200	0.1327120005589860
14	15922.782	15922.782	2005Q2	-0.5219282981215570	0.1841957077683860	0.4926244548107310	0.1871033806300080
15	16047.587	16047.587	2005Q3	-0.5608927886427360	0.0475595077757574	0.7838140345072910	0.1841957077683860
16	16136.734	16136.734	2005Q4	-0.5409926099149390	0.1173431688520470	0.5555165396517350	0.0475595077757574
17	16353.835	16353.835	2006Q1	-0.554769601090578	0.0690315985891669	1.3453837684874700	0.1173431688520470

6. Econometric Modeling

To implement econometric models for predicting GDP growth using sentiment indices.

Steps Taken:

1. **Baseline ARX Model:**
    - An autoregressive model with exogenous predictors (ARX) was used as a baseline.
    - GDP growth rates were regressed on lagged values and macroeconomic factors.
2. **Augmented ARXS Model:**
    - Sentiment indices were added to the ARX model as additional explanatory variables.
3. **Double Lasso Regression:**
    - Implemented to address the issue of variable selection, following the paper's methodology.
    - Steps included:
        1. Identifying predictors of GDP growth using Lasso.
        2. Identifying predictors of sentiment using Lasso.
        3. Combining selected variables into a final OLS model.