BoE - Analytical Approach

Lara Cadel

Table of contents:

1. Notebook 1: Foundation of Data Quality and Initial Insights

The first notebook serves as a cornerstone, emphasizing the importance of robust data preparation as the foundation for meaningful analysis. This draft, the first of four, inspects and assesses the quality of a provided dataset, ensuring its readiness for subsequent analytical phases.

Key Objectives and Workflow

The notebook opens with an introduction to the dataset's metadata, laying out the scope and assumptions regarding its structure and content. This initial understanding is vital for identifying potential gaps or biases. The workflow then follows a systematic path:

1. **Setup and Configurations**
   - Imports necessary libraries and tools for handling data, such as pandas, NumPy, and visualization libraries like Matplotlib and Seaborn.
   - Establishes configurations that streamline subsequent analyses.
2. **Data Ingestion and Inspection**
   - Loads raw data and conducts a preliminary inspection to understand its relevance, completeness, and consistency.
   - Evaluates the format of key variables like economic indicators and text-based components (e.g., speech transcripts).
3. **Data Cleaning and Text Data Preprocessing**
   - Resolves missing values, removes duplicates, and handles outliers.
   - Preprocesses text data by standardizing formats, tokenizing, and removing stopwords.
4. **Exploratory Data Analysis (EDA)**
   - Uses visualizations to identify patterns, trends, and anomalies within the data.
   - Highlights relationships between economic indicators and other **variables.**
5. **Initial Sentiment Analysis**
   - Conducts a preliminary sentiment analysis to validate text data usability.

By the end of this notebook, clean, well-structured data frames are produced, ready for deeper sentiment and correlation analyses in subsequent notebooks.

---

2. Notebook 2: Sentiment Analysis and Correlation Insights

The second notebook builds on the data cleaned and prepared in Notebook 1, focusing on sentiment analysis and its correlation with economic indicators. By leveraging advanced tools like FinBERT, this notebook provides a robust framework for extracting insights into the tone of speeches and its economic implications.

Key Objectives and Methodology

1. **Data Ingestion and Inspection**
   - Loads the cleaned data and reviews metadata for context.
   - Ensures consistency and relevance for sentiment and correlation analyses.
2. **FinBERT Sentiment Analysis**
FinBERT, a pre-trained model fine-tuned for financial contexts, is employed using two approaches:

- **512 Tokens Truncation**: Speeches are segmented into 512-token chunks, with average sentiment scores calculated for each segment. This method is computationally efficient but may lose granularity.
- **FinBERT by Statements**: Analyzes individual statements within speeches, providing a finer-grained view of sentiment. While computationally intensive, it captures nuances missed by the truncation method.

Comparing these methods demonstrates the trade-off between computational efficiency and analytical depth, with statement-level analysis offering deeper insights.

3. Correlation Analysis
- OLS Regression: Explores linear relationships between sentiment scores and economic indicators, identifying statistically significant trends.
- Vector Autoregression (VAR): Captures the dynamic interplay between sentiment and economic indicators over time. This approach accounts for lagged effects, offering a temporal perspective on the relationships.

The insights from this notebook reveal nuanced connections between the tone of central bank speeches and economic trends, serving as a stepping stone for predictive modeling.

---

3. Notebook 3: Predictive Modeling with Sentiment and Economic Indicators

The third notebook ventures into predictive modeling, seeking to quantify the relationship between sentiment scores and economic indicators. It evaluates the feasibility of predicting sentiment trends using advanced statistical and machine-learning models.

Key Objectives and Methodology

1. Random Forest Model
- The Random Forest algorithm is deployed to identify key features influencing sentiment scores.
- Extensive preprocessing, including correlation matrix analysis and Variance Inflation Factor (VIF) calculation, mitigates multicollinearity.
- Despite rigorous feature engineering, the model struggles with accuracy, highlighting the complexity of predicting sentiment from structured economic data.

2. SARIMA Model
- A Seasonal AutoRegressive Integrated Moving Average (SARIMA) model is employed to forecast sentiment trends over time.
- While effective in capturing overarching trends, its predictive precision for short-term variations remains limited. Residual diagnostics validate the model's statistical robustness but underscore its challenges in addressing granular sentiment fluctuations.

This notebook underscores the inherent difficulty in predicting sentiment using traditional models, given the nuanced and context-dependent nature of financial communication. It highlights the need for innovative modeling approaches, potentially incorporating deep learning or alternative statistical techniques.

---

4. Insights from the Notebooks

Together, these notebooks represent a cohesive workflow addressing data preparation, analysis, and prediction.

1. Data Quality as a Pillar
The meticulous cleaning and preprocessing in Notebook 1 establish a solid foundation, ensuring subsequent analyses are built on reliable data.

2. Sentiment Analysis and Correlations
Notebook 2 highlights the value of granular sentiment analysis using FinBERT and the dynamic interplay between speech tone and economic indicators. The dual approach—512 Tokens Truncation and FinBERT by Statements—offers a balanced perspective, catering to both efficiency and depth.

3. Predictive Challenges
Notebook 3 reveals the challenges in modeling sentiment, with traditional methods like Random Forest and SARIMA struggling to capture its complexity. These findings prompt considerations for integrating more sophisticated techniques.

---

5. Future Directions

The project's phased structure leaves room for further exploration:

- Enhanced Models: Future notebooks could experiment with neural networks, particularly transformer-based models like GPT or BERT, for predictive tasks.

- Contextual Analysis: Incorporating additional variables, such as geopolitical events or policy decisions, could enhance the models' explanatory power.

This project's structured approach provides a replicable framework for analyzing the interplay between text data and economic variables, serving as a valuable resource for researchers and practitioners alike.