Homework2 Report

Dongwei Li, Shiqi Ma

1. Web Scraping for FOMC Corpus

In this project, we collected FOMC Meeting Minutes, Fed speeches, and Press Conference transcripts as the corpus for sentiment analysis. Since the Federal Reserve has tended to send signals to the market to effectively regulate it after 2012, we selected texts published from 2012 to the present as our research focus.

Given the large volume of files, we employed web scraping to efficiently collect the textual data. For FOMC Meeting Minutes and Press Conference transcripts, documents prior to 2019 are stored in the historical data pages, while those from 2019 onward are stored in the calendar interface. Therefore, we first directed the scraper to the appropriate storage websites based on the publication year, and then used BeautifulSoup to search for all URLs ending with ".pdf." Based on the naming patterns of the URLs, we filtered out the FOMC Meeting Minutes and Press Conference transcripts and automatically downloaded them. Ultimately, we obtained 101 FOMC Meeting Minutes files. Since Press Conference transcripts were published once every two meetings before 2019, the number of Press Conference transcripts was 76, slightly fewer than the number of FOMC Meeting Minutes.

As for Fed speeches, all documents are stored in the historical data interface. Therefore, we used BeautifulSoup to filter out all URLs ending with ".pdf" on the annual historical data pages and downloaded all files that matched the naming conventions. In the end, we obtained 734 Fed speeches files.

2. Data Preprocessing

After web scraping for FOMC PDF files, we performed data preprocessing on the documents. Initially, we extracted text from each PDF, removing newline characters and hyphenations to ensure word continuity. We then employed regular expressions to split the text into sentences based on punctuation marks like periods and question marks. Subsequently, we filtered these sentences using a list of 15 financial and economic keywords, including "inflation expectation," "interest rate," and "employment." Sentences containing any of these keywords were retained. Each filtered sentence was then saved into a CSV file corresponding to its original document. This method ensures that the processed data is clean, relevant, and systematically organized for further analysis.

3. Sentiment Analysis

(1) Factor Similarity

Factor Similarity is based on the principle of text vectors, analyzing the semantic similarity between two sentences by calculating the distance between their vectors. In this project, the focus of the FOMC texts is on interest rates and inflation. We selected the sentences "Interest rates will rise" and "Inflation will rise" as reference sentences for the Federal Reserve's stance on interest rates and inflation, respectively. Specifically, we used the sentence factor method to convert each sentence in the text into a vector, and then calculated the cosine similarity between this vector and the vectors of the two reference sentences. This cosine similarity serves as the sentiment index of each sentence toward interest rates and inflation. By averaging the similarities of all sentences in a document, we obtained the document’s average sentiment toward interest rates and inflation, recorded as the int score and inf score, respectively. In this way, we derived two polarity indicators, which reflect the Federal Reserve’s sentiment toward interest rates and inflation.

(2) RoBERTa

We also utilize a pre-trained RoBERTa model[[1]](#footnote-1) to discern the Federal Reserve's sentiment regarding economic conditions. This model, which is adept at classifying sentiments, categorizes each sentence from the documents into three sentiment states: hawkish, dovish, or neutral. Specifically, our method involves an initial step where each sentence is analyzed and classified according to these sentiment categories. Following this classification, we compute a sentiment index for each document. This index is calculated as the net sentiment score, representing the normalized difference between the counts of hawkish and dovish statements. This score quantitatively captures the overall sentiment direction of the document, providing a clear indicator of the Federal Reserve's leaning towards either a tightening or loosening of monetary policy. This systematic approach allows us to aggregate sentiment data across multiple documents, offering a nuanced view of the Federal Reserve's economic stance over time.

4. Regression Analysis

(1) Factor Similarity

We conducted a regression analysis to study the impact of the int score and inf score from FOMC Meeting Minutes, Fed speeches, and Press Conference transcripts on the daily changes in 10-year Treasury yield, 2-year Treasury yield, the spread between 10-year and 2-year Treasury yields, gold prices, the S&P 500, and the VIX. The regression results are shown in the table. From the results, we can see that the int score and inf score of the FOMC Meeting Minutes exhibit the strongest significance, followed by Fed speeches, while Press Conference transcripts are not significant. This indicates that FOMC Meeting Minutes convey the most information to the market, Fed speeches also contain relevant information, whereas Press Conference transcripts carry little to no information.

In terms of asset categories, the regression results show the strongest significance for the 10-year Treasury yield, the 2-year Treasury yield, and gold prices. This suggests that the Federal Reserve's sentiment has the most pronounced impact on assets that are directly linked to interest rates and monetary policy, such as Treasury bonds and gold.

Specifically, the regression coefficient of the int score on changes in Treasury yields is significantly negative, which may indicate a reversal effect after the market overreacts to the Federal Reserve's monetary policy. Similarly, the regression coefficient of the inf score on changes in Treasury yields is also significantly negative. This could be because when the market hears statements from the Federal Reserve regarding rising interest rates or inflation, some investors choose to move toward safe assets like Treasury bonds. This flight to safety behavior drives up the price of Treasury bonds, thereby lowering their yields.

(2) RoBERTa

We also conducted a regression analysis between the Sentiment Index from RoBERTa and key market indicators. The results show that Press Conference transcripts have the most significant influence, particularly on the 10-year Treasury yield and the S&P 500, with strong positive and statistically significant relationships. This suggests that sentiment from Press Conferences plays a critical role in shaping market reactions, especially for long-term interest rates and equity performance.

FOMC Meeting Minutes also exhibit some significant effects, particularly on gold prices and the Treasury yield spread. For example, the positive coefficient on gold prices suggests that dovish sentiment in meeting minutes pushes investors towards safe-haven assets like gold. However, the influence of Meeting Minutes is generally less pronounced than that of Press Conferences.

In contrast, Fed speeches have the weakest impact, with only sporadic significance, primarily on the S&P 500. This indicates that speeches carry less market-moving information compared to other forms of Federal Reserve communication.

Appendix

Table 1 Regression Result of Factor Similarity for FOMC Minutes

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **GT10** | | **GT2** | | **GTspread** | | **Gold** | | **SP500** | | **VIX** | |
|  | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** |
| **constant** | 1.1866\*\*\* | 25.6282 | 1.3447\*\*\* | 49.9109 | -16.0634 | -141.1604 | -164.9783 | -9.1947 | 196.1680 | 6.6501 | 12.6934 | 62.6661 |
|  | (0.4437) | (19.9772) | (0.4575) | (48.4388) | (29.7652) | (113.1398) | (115.1634) | (7.2870) | (246.1316) | (7.2217) | (13.0858) | (62.5438) |
| **int\_score** | -3.7425\*\* | -91.1372 | -3.9441\*\* | -236.0855 | 23.6091 | 784.8858 | 1017.4270\*\* | 58.8272\* | -989.6168 | -31.3374 | 23.2275 | 197.7944 |
|  | (1.8557) | (83.5557) | (1.9137) | (202.5984) | (124.4947) | (473.2141) | (481.6779) | (30.4785) | (1029.4606) | (30.2051) | (54.7321) | (261.5934) |
| **inf\_score** | 1.3708 | 39.3952 | 1.2647 | 136.5459 | 7.6791 | -504.1080 | -680.4851\* | -40.0235 | 611.0504 | 18.4896 | -48.2567 | -320.4229 |
|  | (1.5291) | (68.8506) | (1.5769) | (166.9427) | (102.5846) | (389.9321) | (396.9063) | (25.1145) | (848.2836) | (24.8892) | (45.0997) | (215.5550) |
| **R²** | 0.0738 | 0.0193 | 0.0857 | 0.0159 | 0.0032 | 0.0290 | 0.0450 | 0.0376 | 0.0102 | 0.0124 | 0.0243 | 0.0364 |

Standard errors in parentheses. \*p < 0.10," "\*\*p < 0.05," "\*\*\*p < 0.01" are used to denote significance at the 10%, 5%, and 1% levels, respectively.

Table 2 Regression Result of Factor Similarity for FOMC Press Conference Transcripts

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **GT10** | | **GT2** | | **GTspread** | | **Gold** | | **SP500** | | **VIX** | |
|  | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** |
| **constant** | 0.0471 | -28.7118 | -0.0623 | -43.4892 | 10.5851 | 175.2972 | 223.1003 | 12.4163 | 288.7497 | 8.1661 | 9.1182 | 27.4190 |
|  | (0.5735) | (26.5907) | (0.6268) | (60.4946) | (35.5695) | (175.1974) | (145.4950) | (9.0560) | (328.2567) | (9.2578) | (16.9920) | (78.5722) |
| **int\_score** | 0.3380 | 37.5120 | 0.6333 | 58.0934 | -28.0965 | -678.2983 | -209.9727 | -10.1495 | -1066.3420 | -23.1025 | 3.7720 | -10.1069 |
|  | (1.4522) | (67.3296) | (1.5871) | (153.1765) | (90.0645) | (443.6121) | (368.4034) | (22.9304) | (831.1688) | (23.4414) | (43.0250) | (198.9503) |
| **inf\_score** | -0.4473 | 15.1708 | -0.5285 | 22.5116 | 7.4443 | 317.8305 | -193.8811 | -12.1743 | 483.2424 | 7.1881 | -20.7799 | -44.2216 |
|  | (0.9421) | (43.6823) | (1.0297) | (99.3782) | (58.4323) | (287.8077) | (239.0136) | (14.8768) | (539.2476) | (15.2084) | (27.9139) | (129.0754) |
| **R²** | 0.0032 | 0.0159 | 0.0038 | 0.0072 | 0.0015 | 0.0318 | 0.0362 | 0.0307 | 0.0225 | 0.0139 | 0.0112 | 0.0034 |

Standard errors in parentheses. \*p < 0.10," "\*\*p < 0.05," "\*\*\*p < 0.01" are used to denote significance at the 10%, 5%, and 1% levels, respectively.

Table 3 Regression Result of Factor Similarity for FOMC Speeches

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **GT10** | | **GT2** | | **GTspread** | | **Gold** | | **SP500** | | **VIX** | |
|  | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** |
| **constant** | 0.0471 | 2.7434\*\* | 0.0185 | 3.2432 | 2.8663 | -1.9836 | -15.9299\* | -0.9909\* | -18.1204 | -0.1462 | -0.2610 | -0.2334 |
|  | (0.0296) | (1.3021) | (0.0266) | (2.2431) | (2.0543) | (18.4764) | (8.7738) | (0.5214) | (17.0206) | (0.4875) | (0.7674) | (3.9700) |
| **int\_score** | 0.1003 | 3.3872 | 0.1068 | -2.7943 | -0.6193 | 72.8294 | 8.9534 | 0.2168 | 30.4518 | -0.4319 | 0.8059 | 7.8832 |
|  | (0.1312) | (5.7661) | (0.1178) | (9.9334) | (9.0971) | (81.8201) | (38.8535) | (2.3090) | (75.3733) | (2.1589) | (3.3982) | (17.5805) |
| **inf\_score** | -0.1876\* | -8.5420\* | -0.1384 | -3.1111 | -4.9525 | -67.2295 | 23.9113 | 1.7987 | 10.1355 | 0.8360 | -0.3161 | -7.2301 |
|  | (0.1111) | (4.8795) | (0.0997) | (8.4060) | (7.6983) | (69.2394) | (32.8794) | (1.9540) | (63.7839) | (1.8270) | (2.8757) | (14.8774) |
| **R²** | 0.0082 | 0.0113 | 0.0037 | 0.0029 | 0.0036 | 0.0014 | 0.0063 | 0.0072 | 0.0022 | 0.0006 | 0.0002 | 0.0004 |

Standard errors in parentheses. \*p < 0.10," "\*\*p < 0.05," "\*\*\*p < 0.01" are used to denote significance at the 10%, 5%, and 1% levels, respectively.

Table 4 Regression Result of RoBERTa for FOMC Minutes

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **GT10** | | **GT2** | | **GTspread** | | **Gold** | | **SP500** | | **VIX** | |
|  | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** |
| **constant** | -0.0196\*\*\* | -0.6439\*\* | -0.0191\*\*\* | -0.3714 | -0.0454 | 0.2414 | 4.3501\*\* | 0.2464\*\* | -0.2986 | -1.3313 | 5.3438 | 0.1634 |
|  | (0.0067) | (0.2985) | (0.007) | (0.7227) | (0.4414) | (1.6652) | (1.7425) | (0.1099) | (0.193) | (0.9326) | (3.6525) | (0.1076) |
| **Sentiment Index** | -0.0526\* | -0.5488 | -0.0580\* | 1.2019 | 0.5569 | 15.5718\*\* | 4.6266 | 0.2336 | -1.5906\* | -6.4957 | 13.3289 | 0.1011 |
|  | (0.0307) | (1.3601) | (0.0318) | (3.2928) | (2.011) | (7.5865) | (7.9388) | (0.5007) | (0.8794) | (4.2491) | (16.6406) | (0.4903) |
| **R²** | 0.0292 | 0.0017 | 0.0329 | 0.0014 | 0.0008 | 0.0412 | 0.0035 | 0.0022 | 0.0323 | 0.0233 | 0.0065 | 0.0004 |

Standard errors in parentheses. \*p < 0.10," "\*\*p < 0.05," "\*\*\*p < 0.01" are used to denote significance at the 10%, 5%, and 1% levels, respectively.

Table 5 Regression Result of RoBERTa for FOMC Press Conference Transcripts

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **GT10** | | **GT2** | | **GTspread** | | **Gold** | | **SP500** | | **VIX** | |
|  | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** |
| **constant** | -0.0257\*\*\* | -0.9572\*\* | -0.0279\*\*\* | -1.1098 | 0.2295 | 1.5594 | 7.4203\*\*\* | 0.4385\*\*\* | -0.4478\* | -2.2807\* | 7.0395 | 0.2499\* |
|  | (0.0084) | (0.3995) | (0.0092) | (0.9032) | (0.5314) | (2.5896) | (2.2124) | (0.1373) | (0.2546) | (1.1743) | (4.9563) | (0.1386) |
| **Sentiment Index** | -0.0665\* | -0.8163 | -0.0673\* | 2.7734 | 0.0786 | 21.9470\* | -0.5201 | -0.0966 | -0.5708 | -1.4837 | -1.4895 | -0.4478 |
|  | (0.0359) | (1.7129) | (0.0394) | (3.8721) | (2.2781) | (11.1018) | (9.4846) | (0.5886) | (1.0916) | (5.0341) | (21.2476) | (0.5944) |
| **R²** | 0.0448 | 0.0031 | 0.0384 | 0.007 | 0 | 0.0508 | 0 | 0.0004 | 0.0037 | 0.0012 | 0.0001 | 0.0077 |

Standard errors in parentheses. \*p < 0.10," "\*\*p < 0.05," "\*\*\*p < 0.01" are used to denote significance at the 10%, 5%, and 1% levels, respectively.

Table 6 Regression Result of RoBERTa for FOMC Speeches

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **GT10** | | **GT2** | | **GTspread** | | **Gold** | | **SP500** | | **VIX** | |
|  | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** | **\_change** | **\_%change** |
| **constant** | 0.0023 | 0.1555 | 0.0017 | 0.3021 | 0.0637 | 0.4217 | 0.4633 | 0.0206 | -0.0664 | -0.2106 | 2.5703\* | 0.0860\*\* |
|  | (0.0025) | (0.1128) | -0.0022 | -0.2008 | -0.1669 | -1.8304 | -0.7185 | -0.0439 | -0.064 | -0.3396 | -1.4945 | -0.0434 |
| **Sentiment Index** | 0.0134 | 0.1619 | 0.0127 | -0.78 | 0.0473 | -9.2459 | -1.9244 | -0.1633 | 0.1075 | 0.7229 | 4.9849 | 0.0177 |
|  | (0.0108) | (0.4819) | (0.0094) | (0.8577) | (0.7132) | (7.8199) | (3.0695) | (0.1873) | (0.2748) | (1.4577) | (6.4272) | (0.1866) |
| **R²** | 0.0035 | 0.0003 | 0.0042 | 0.0019 | 0 | 0.0032 | 0.0009 | 0.0018 | 0.0004 | 0.0006 | 0.0014 | 0 |

Standard errors in parentheses. \*p < 0.10," "\*\*p < 0.05," "\*\*\*p < 0.01" are used to denote significance at the 10%, 5%, and 1% levels, respectively.

1. https://github.com/gtfintechlab/fomc-hawkish-dovish [↑](#footnote-ref-1)