Introduction

In this project, we analyzed the hawkishness and dovishness of the Federal Open Market Committee (FOMC) communications, including Meeting Minutes, Fed Speeches, and Press Conference Transcripts from 2012 to 2024. We conducted sentiment analysis using Factor Similarity Analysis and the FOMC RoBERTa model developed by Georgia Tech to categorize the texts of public annoucements as hawkish, dovish, or neutral. Building on this classification, we examined the relationship between key financial indicators—such as the 2y-10y Treasury yield spread—and the sentiment expressed by the Fed. We explored how the Fed's decisions could potentially trigger a "butterfly effect" in financial market using real data.

Method 1

The first method in this project is the Factor Similarity Method that projects sentences into high-dimensional vectors, allowing for the quantification of their relevance through the cosine similarity formula: $cos\theta = \frac{A \cdot B}{\|A\| \|B\|}$. We applied a pre-trained FinBERT model, utilizing it as a transfer learning technique.

FinBERT's architecture incorporates multiple attention layers, and a special classification token (**[CLS]**) at the beginning of each input. This token captures a summarized representation of the entire input sentence or document via attention mechanism. The attention mechanism allows FinBERT to focus on different contextual parts of the text and aggregate this information into the **[CLS]** token (weighted vector). This approach ensures the generation of fixed length vectors with context understanding for every document released by FOMC and allows us to evaluate and construct hawkish and dovish methods.

```
A1 = ["inflation expectation", "interest rate", "bank rate", "fund rate", "price", "economic activity", "inflation",
      "employment"]

B1 = ["unemployment", "growth", "exchange rate", "productivity", "deficit", "demand", "job market", "monetary policy"]
```

(Figure 1.1: Dictionary filters adapted from FOMC RoBERTa model)

One important design aspect worth mentioning is that our team implemented a dictionary filter to address the issue of noisy financial data, as discussed in the article *Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis*. Including irrelevant words can introduce biases, so like the authors, we applied a dictionary filter that only retains words found in the **A1** and **B1** sections, as outlined in the linked code above.

We have also defined our **query vectors** to be as detailed as possible. For example, the hawkish sentence is: "To prevent rising inflation, the central bank will raise interest rates to maintain price stability," while the dovish sentence is: "Inflation is expected to rise, but the central bank will maintain low interest rates to support economic growth." By designing these query vectors and

filtering out irrelevant words carefully, we aim to boost our analysis accuracy and overall effectiveness in identifying sentiment patterns.

Using these methodologies assign average dovish and hawkish similarities for each document each day, average them and make them as independent variables that regress against intra day 2Y change, intra day 10Y change, and **next day change of 10Y-2Y spread.** The output is shown as following:

OLS Regression Results

Dep. Variable:	Change	R-squared:	0.001
Model:	OLS	Adj. R-squared:	-0.002
Method:	Least Squares	F-statistic:	0.4728
Date:	Sat, 12 Oct 2024	Prob (F-statistic):	0.623
Time:	09:11:18	Log-Likelihood:	1061.7
No. Observations:	683	AIC:	-2117.
Df Residuals:	680	BIC:	-2104.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0320	0.032	-1.005	0.315	-0.095	0.031
dovish_similarity	0.0080	0.088	0.091	0.928	-0.165	0.181
hawkish_similarity	0.0456	0.087	0.526	0.599	-0.124	0.216

Omnibus:	138.978	Durbin-Watson:	2.146
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1187.151
Skew:	-0.642	Prob(JB):	1.64e-258
Kurtosis:	9.330	Cond. No.	75.8

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

(Figure 1.2: 2 Year Linear Regression Output)

OLS Regression Results

Dep. Variable:	Change	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.003
Method:	Least Squares	F-statistic:	0.04753
Date:	Sat, 12 Oct 2024	Prob (F-statistic):	0.954
Time:	09:11:19	Log-Likelihood:	1004.3
No. Observations:	683	AIC:	-2003.
Df Residuals:	680	BIC:	-1989.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0115	0.035	-0.331	0.741	-0.080	0.057
dovish_similarity	0.0040	0.096	0.041	0.967	-0.184	0.192
hawkish_similarity	0.0147	0.094	0.156	0.876	-0.170	0.200

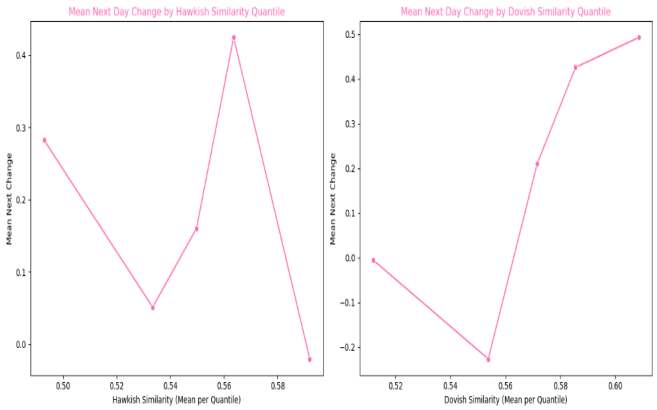
Omnibus:	18.435	Durbin-Watson:	2.046
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31.661
Skew:	-0.180	Prob(JB):	1.33e-07
Kurtosis:	3.991	Cond. No.	75.8

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Same Happens with Long

(Figure 1.3: 10 Year Linear Regression Output)

Both 2Y and 10Y regression output suggest that intra-day changes do not fully capture the market's response to the hawkish or dovish sentiment in FOMC communications. This is reflected in both the low statistical significance and r-squared. Nevertheless, this outcome may be influenced by the fact that many announcements **are made in the afternoon (verified by professor)**, which limits the market's ability to respond immediately. We also hypothesize that the yield spread might offer a clearer reflection of market sentiment, as both the 10-year and 2-year yields can rise simultaneously, irrespective of whether the announcement is hawkish or dovish. On the other hand, the spread tends to widen when the sentiment is dovish, reflecting market expectations of lower rates and longer-term economic support.



(Figure 1.4: Similarity Score Quantile vs Aggregated mean of next day spread change)

OLS Regression Results

Dep. Variable:	next_change	R-squared:	0.009
Model:	OLS	Adj. R-squared:	0.006
Method:	Least Squares	F-statistic:	3.132
Date:	Sat, 12 Oct 2024	Prob (F-statistic):	0.0443
Time:	09:11:17	Log-Likelihood:	-1848.3
No. Observations:	683	AIC:	3703.
Df Residuals:	680	BIC:	3716.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.5094	2.259	-1.111	0.267	-6.945	1.926
dovish_similarity	15.5604	6.241	2.493	0.013	3.306	27.815
hawkish_similarity	-11.2055	6.134	-1.827	0.068	-23.250	0.839

Omnibus:	35.910	Durbin-Watson:	1.804
Prob(Omnibus):	0.000	Jarque-Bera (JB):	87.573
Skew:	0.252	Prob(JB):	9.63e-20
Kurtosis:	4.680	Cond. No.	75.8

(Figure 1.5: Regression analysis for 10-2 yr spread)

As expected, the 10-2 year spread change on the next date shows statistical significance for both dovish and hawkish factors. We observe that, on average, a 0.1 percent increase in dovish sentiment tends to widen the spread by 1.55 units, while a 0.1 percent increase in hawkish sentiment is associated with a decrease in the spread by 11.2055 units. However, the R-squared seems very low which may have indicated lack of linear relationship or there lacks control variables to explain variances of the spread patterns

In conclusion, the similarity score shows no explanatory ability for the GY10 and GY2 yields, as reflected by the insignificant patterns observed. However, it does **have some explanatory power** for the spread, given that spreads are typically more sensitive to hawkish or dovish tones in monetary policy. The extremely small R-squared value suggests that the similarity score either lacks a strong relationship with the spread or that there are missing control variables that could better explain the variances in spread movements.

Despite these limitations, the similarity score may still provide value by helping to refine the data more precisely. This could involve creating a better filter dictionary, applying more exact regex methods, or defining more accurate "query" sentences for cosine similarity calculations.

Nevertheless, we doubt that the similarity scores alone serve as a prominent factor in explaining market dynamics. They may offer insights into the tone of central bank communications but should never be relied upon for trading decisions. The scores can support analysis but not act as standalone predictors in financial models.

Method 2

We use the FOMC-RoBERTa model to classify sentences based on their hawkish or dovish tone. This pretrained model comes from the paper "Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis." By using their dictionary to filter out relevant sentences for further classification.

Our regressor in following regression is the document-level hawkishness, computed with the formula in the paper:

$$Measure_i = \frac{\#Hawkish_i - \#Dovish_i}{\#Total_i}$$

After preprocessing, we obtained 716 effective observations for the level of hawkishness, with 182 of them measured as neutral.

Typically, when hawkish sentiment occurs, the yield spread narrows (flattening or even inverting), while dovish sentiment tends to widen the spread. In our analysis, a higher

hawkishness measure value indicates more hawkish sentiment, and we would expect the yield spread to decrease.

The regression results for the level of hawkishness and its impact on the 10-2 year spread, 2-year yield, and 10-year yield all show a lack of statistical significance, suggesting that hawkishness alone may not be a strong predictor of yield changes. Among three regression results, the regression on the 10-2 year spread spread has the lowest p-value of 0.128 and a coefficient of -0.9951, which aligns with expectations that a more hawkish stance tends to narrow the spread. The graph on the left indicates a negative relationship between hawkishness and the 10-2 year spread, as more hawkish speeches are associated with a narrowing yield curve. This reflects market expectations of more aggressive monetary tightening, particularly in the short term. However, the mean spread change is more volatile when the level of hawkishness is between -0.1 and 0.1.

OLS Regression Results

Dep. Variable:	next_change	R-squared:	0.003
Model:	OLS	Adj. R-squared:	0.002
Method:	Least Squares	F-statistic:	2.317
Date:	Sat, 12 Oct 2024	Prob (F-statistic):	0.128
Time:	21:28:44	Log-Likelihood:	-1850.2
No. Observations:	683	AIC:	3704.
Df Residuals:	681	BIC:	3714.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1671	0.139	1.198	0.231	-0.107	0.441
Hawkish Measure	-0.9951	0.654	-1.522	0.128	-2.279	0.288

Omnibus:	37.587	Durbin-Watson:	1.801
Prob(Omnibus):	0.000	Jarque-Bera (JB):	86.254
Skew:	0.294	Prob(JB):	1.86e-19
Kurtosis:	4.638	Cond. No.	4.70

Figure 2.1: 10y-2y Spread Regression

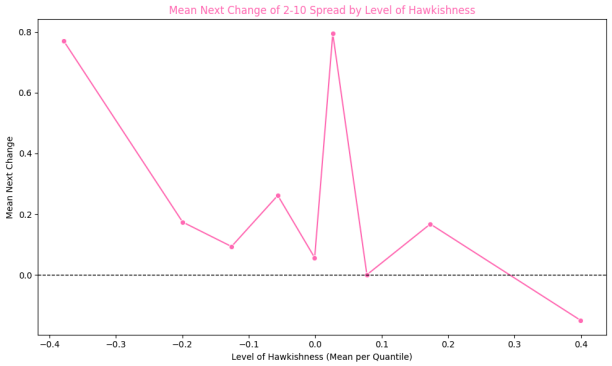


Figure 2.2: Hawkish Measure Quantile vs NextChange 10y-2y Spread

As factor analysis performed better when using both average dovish and hawkish similarities as regressors, we attempted to separate the hawkish measure into two independent variables: one for hawkish and one for dovish tones. The results, shown in Figure 2.3, indicate that positive values represent a hawkish tone, while negative values represent a dovish tone. Both variables yielded negative coefficients with large p-values, suggesting that separating them was not an effective approach. The regression using only the hawkish measure did not produce statistically significant results in our analysis, unlike the findings in the paper. This discrepancy may be due to differences in model fine-tuning or data processing techniques.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            next_change   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     1.370
Date:                Sun, 13 Oct 2024   Prob (F-statistic):              0.255
Time:                        00:09:04   Log-Likelihood:                -1850.0
No. Observations:                 683   AIC:                             3706.
Df Residuals:                     680   BIC:                             3720.
Df Model:                           2
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                            0.0830      0.190      0.437      0.662      -0.290       0.456
Hawkish Measure Positive        -0.4035      1.118     -0.361      0.718      -2.599       1.792
Hawkish Measure Negative        -1.5651      1.091     -1.434      0.152      -3.708       0.578
==============================================================================
Omnibus:                       36.902   Durbin-Watson:                   1.798
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               84.807
Skew:                           0.287   Prob(JB):                     3.84e-19
Kurtosis:                       4.628   Cond. No.                         9.14
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 2.3: 2y-10y Spread Regression with Hawkish and Dovish Separated

For the 2-year yield, the Figure 2.3 suggests that super hawkish speeches will lead to an increase in average short-term rates, and dovish documents tend to drive the interest rate down with a smaller effect. However, the relationship between hawkishness and rate changes is not linear. The yield change is also much more volatile when the tone is between neutral and light hawkish (0 to 0.2). Similarly, for the 10-year yield, the most hawkish speeches (around 0.4) suggests market anticipation of prolonged tightening measures. However, none of these effects are captured significantly in the regression models, highlighting potential limitations in using hawkishness alone as a predictor of yield changes.

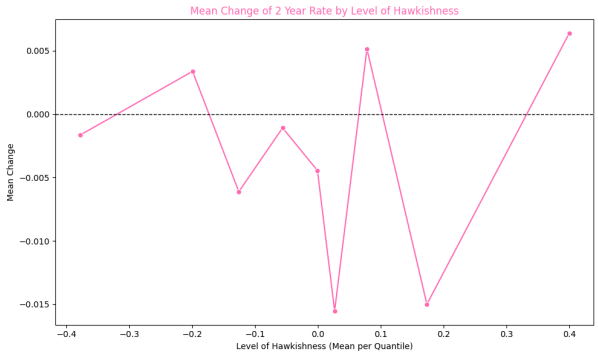


Figure 2.4: Hawkish Measure Quantile vs Mean 2-year rate Change

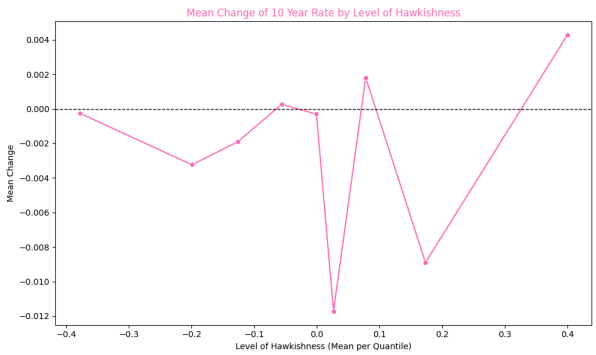


Figure 2.5: Hawkish Measure Quantile vs Mean 10-year rate Change

Comparison and Conclusion

Based on our discoveries regarding the 10-year to 2-year spread, it appears that the cosine similarity method yields better results. We have not observed statistically significant outcomes when applying the second method, which uses the FOMC-RoBERTa (A model Fine Tuned on FOMC) model. This difference probably yields different word comprehension mechanisms for both models.

The cosine similarity method measures the overall relevance of a sentence to predefined hawkish or dovish vectors (the more detailed the better), capturing more subtle sentiment expressions. It allows for a more contextually aware interpretation of sentences, detecting hawkish or dovish tones even when specific keywords are not present and does not require model training.

On the other hand, the FOMC-RoBERTa model relies on classification based on carefully annotated sentences provided by the authors of the "Trillion Dollar Words" paper. These annotations were used to build a labeled dataset, which includes sentences specifically identified as hawkish or dovish (1,070-sentence Meeting Minutes dataset, a 315-sentence Press Conference dataset, and a 994-sentence Speech dataset).

While these datasets are useful for training the model, the reliance on this specific sampling may limit the model's ability to capture subtler sentiments or expressions not covered in the original annotations (especially when we have different data processing processes). For instance, a statement indicating caution without explicitly stating "hawkish" or "dovish" terms might still influence market movements. This issue could be further compounded by the introduction of an additional 'neutral' class in the FOMC-RoBERTa model, as unseen data might be incorrectly classified as neutral. The reliance on a predefined training set may cause the model to misclassify sentences that don't closely align with the training data into the neutral class, potentially reducing the accuracy of hawkish or dovish sentiment detection.

In that fashion, Cosine similarity is more ROBUST as they can detect such indirect cues, whereas a keyword-based filter might overlook them if they are not present in the limited training set it has seen, which is exactly why we saw such a phenomenon in our analysis.

One important consideration is that a word-list-based method could also be explored in this project as an alternative for sentiment analysis. This approach would involve manually constructing lists of hawkish and dovish terms and using those lists to classify sentences. However, we doubt this method would perform well without significant effort and expertise. For example, building an effective word list requires carefully selecting terms that accurately reflect hawkish or dovish sentiment across a wide range of contexts, similar to how we designed query vectors for cosine similarity (Professor mentioned it will update every year and last for five years).

Given our limited time and lack of deep knowledge in fixed-income markets, this method could lead to inaccuracies, as it might overlook complex expressions. For instance, a term like "rate increase" might not always signal hawkish sentiment if the context implies future cuts. Constructing such a word list without sufficient expertise could lead to misclassification or overly simplistic results.

Therefore, we believe that approaches like FinBERT with cosine similarity, which use high-dimensional vectors to capture the relationships between words and sentences, would likely be more effective for analysts without strong domain knowledge. These methods allow for greater flexibility in understanding financial language, avoids the need for manually curated word lists and can better account for subtle shifts in sentiment. As a result, we will make it the most preferable approach, with finbert sentimental comes next, and word list comes last.