

Sentiment-Model Complexity in Low-Frequency Financial-Text Environments

Andreas Stylianos Metaxas

Department of International and European Economic Studies

Thesis supervisor: Vasia Panousi

Abstract

This paper examines whether more complex NLP methods improve the usefulness of sentiment signals extracted from low-frequency financial text. Using FOMC minutes from 2000 to 2025, the analysis compares the Loughran and McDonald financial sentiment lexicon, FinBERT, and a Masked Language Modeling adaptation, evaluating each against five-day changes in the VIX using OLS and GARCH-based specifications. The results show that greater NLP complexity does not necessarily lead to better empirical performance. FinBERT produces the strongest in-sample association with VIX changes, especially under the GARCH framework, while the lexicon-based approach remains a competitive and interpretable benchmark. By contrast, the MLM-adapted model performs worse, suggesting that excessive domain adaptation may reduce economically useful variation in the sentiment signal. Overall, the findings indicate that, in low-frequency settings, model complexity is useful only when it improves the match between the sentiment signal, the data structure, and the econometric framework.

1 Introduction

Analysis of financial markets relies on textual data to extract signals about economic conditions and expectations that shape market behavior. However, a significant problem arises when only low-frequency data are available, which, unlike high-frequency data that allows models to learn patterns, provides limited data points that naturally increase the risk of overfitting and reduce reliability. This raises an important question about the relationship between complexity and predictive power in similar environments. In this paper, FOMC minutes are used to provide a neutral setting to study this issue, as they are released infrequently and follow structured language.

The limitations posed by low-frequency data become increasingly important when selecting the appropriate modeling approach. In environments that require low-frequency data, the benefits of complex models are not certain, as the number of data points may not provide the necessary information for the model to extract meaningful patterns. Traditional Lexicon-Based approaches, such as the LM lexicon, offer a simple baseline that relies on a set of predefined dictionaries to extract sentiment without requiring high amounts of computational power. In contrast, transformer-based models provide a more in-depth look at the data by incorporating contextual awareness and understanding semantic relationships, at the cost of increased computational expense. While both approaches are widely used in the literature, it remains unclear whether the benefits of a more complex model outweigh the increased costs, even when paired with the correct econometric model.

This paper contributes to the literature by assessing the relationship between complexity and predictive power in low-frequency financial environments. It compares the Loughran and McDonald financial sentiment lexicon method with transformer-based models, including FinBERT and a domain-specialized adaptation through Masked Language Modeling (MLM). To fully understand this relationship, we use both linear and nonlinear approaches within the appropriate econometric framework to assess the effectiveness of the sentiment signals. The major contribution of this paper is the finding that performance degrades with increased specificity, indicating that, with these kinds of systems, complexity does not necessarily mean better performance.

To answer this question, the paper uses a dataset composed of FOMC minutes covering the period 2000–2025 and uses three main approaches to extract the sentiment signal, as stated before: the Loughran-McDonald Lexicon, pre-trained FinBERT, and Masked Language Modeling, along with a Low-Rank Adaptation (LoRA) based variation to examine whether low-parameter fine-tuning can preserve performance while reducing costs. To preserve order within the dataset and correctly align with the implied volatility of the market, which is a daily observation, the data are aligned using an "as-of" procedure that matches each document to the next valid trading day. To evaluate the sentiment, both a linear and a nonlinear approach are used, more specifically, Ordinary Least Squares (OLS) and an event-driven GARCH specification, to assess their explanatory performance in the low-frequency setting established.

The results suggest a trade-off between complexity and predictive power. On the one hand, the complex NLP models output rich sentiment signals, but these gains from complexity do not always translate into better performance. Specifically, FinBERT produces the

most robust and strongest in-sample fit, and, when combined with the GARCH framework, it outputs the most statistically significant relationship with VIX. On the other hand, the LM-lexicon remains a useful interpretable benchmark, which performs on par with FinBERT out of sample, suggesting that simplicity can still be a valuable tool in low-frequency environments. In contrast, the MLM variant has the worst results, which suggests that over-specificity can hinder predictive performance. Finally, the results indicate that additional complexity in the econometric layer is justified when it addresses clear features of the data, particularly conditional heteroskedasticity in volatility-related outcomes. Overall, the evidence suggests that in the NLP layer there is a trade-off between complexity and predictive power, and in the econometric layer, a well-fitted model, complex or not, is necessary.

The findings discussed above matter because they suggest that complexity should not be the sole criterion in sentiment analysis for low-frequency data with limited data points. As shown, greater complexity does not necessarily lead to better results, and overspecialization, as in domain-specific adaptation, may even reduce performance. On the other hand, results show that a well-fitted econometric model has a crucial role in order to translate sentiment into inference. The paper argues that the optimal point in a low-frequency environment does not lie in maximizing complexity at each layer but rather in combining a suitable generalized model and picking a framework that fits the data.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature on financial sentiment, model complexity and econometric models. Section 3 describes the data. Section 4 describes the NLP and econometric methodologies used. Section 5 presents and interprets the results. Section 6 discusses robustness tests. Section 7 concludes.

2 Related Literature

The use of textual data in financial economics has expanded substantially, as researchers have sought to extract market-relevant information from corporate disclosures, news articles, and central-bank communications. A crucial methodological issue in this literature is that financial language differs from ordinary language. Loughran and McDonald[12] show that many words classified as negative by general-purpose dictionaries do not carry the same meaning in financial contexts, motivating the creation of their financial dictionary. Tetlock[16] provides early evidence that textual pessimism in financial news contains information relevant for stock prices and trading activity. These studies establish that financial text may contain economically meaningful signals, but they also suggest that the construction of the sentiment measure is essential. The present thesis builds on this literature by treating the Loughran-McDonald dictionary as a simple, interpretable, and domain-specific baseline against which more complex sentiment methods can be evaluated.

More recent work has shifted the focus from dictionary-based sentiment measurement to contextual language models. Araci[3] introduces FinBERT, a transformer-based model adapted to financial text, and shows that it improves performance on financial sentiment classification tasks relative to earlier machine-learning approaches. This development is important because transformer models can account for context, syntax, and semantic relationships that bag-of-words methods ignore. Apel and Grimaldi[1] also emphasize that

central-bank minutes contain richer information than static word-count approaches may be able to capture. This literature suggests that more complex language models should, in principle, extract a more informative sentiment signal from financial and monetary policy documents. However, this expectation is less straightforward when the text is released infrequently, and the object of interest is a noisy, high-frequency market variable. The present thesis therefore does not frame the comparison as one between an outdated method and a superior modern method, but rather as a comparison between simplicity and contextual complexity under severe sample constraints.

The monetary-policy communication literature provides the institutional setting for this comparison. Gurkaynak, Sack, and Swanson[7] show that asset-price responses to Federal Open Market Committee (FOMC) events cannot be explained only by the current target-rate surprise, since communication also conveys information about the expected future path of policy. Lucca and Trebbi[14] show that automated text methods can quantify FOMC communication and that this communication component helps explain Treasury yields and future policy movements. Hansen, McMahon, and Prat[9] use computational linguistics to study deliberation within the FOMC, demonstrating that central-bank text contains measurable variation in topics and information. Hubert and Labondance[11] find that the tone of central-bank statements explains monetary-policy surprises beyond other same-day information, while Tadle[15] shows that sentiment extracted from FOMC documents is associated with financial-market variables such as fed funds futures and the US dollar. These studies support the view that central-bank communication contains market-relevant information. In this thesis, the focus is on the FOMC minutes, as it is a structured and regularly scheduled low-frequency disclosure setting to analyze whether more sophisticated textual models generate information that is economically more informative.

Related literature examines the effect of monetary-policy news on financial-market volatility. Bomfim[5] finds that surprise monetary-policy decisions increase stock-market volatility, and that the volatility response differs depending on the sign of the policy surprise. Apergis[2] uses GARCH modeling and finds that FOMC minutes affect the mean and the volatility of asset prices, particularly before the global financial crisis. Lucca and Moench[13] document a predictable equity-price drift before scheduled FOMC announcements. These papers show that monetary-policy communication is relevant not only for asset-price levels but also for uncertainty and volatility. At the same time, they suggest that the market response to FOMC communication is event-driven and potentially regime-dependent. This is important for the present thesis because the dependent variable, changes in the VIX, is observed at high frequency and is influenced by many sources of market news beyond FOMC minutes. The empirical problem is therefore characterized by a frequency mismatch: infrequent textual disclosures are used to explain movements in a noisy, high-frequency measure of market uncertainty.

The econometric literature on volatility further motivates the modeling strategy used in this thesis. Bollerslev[4] introduces the GARCH framework to model conditional heteroskedasticity and volatility clustering, features central to financial time series. Engle and Ng[6] show that volatility can react asymmetrically to news, implying that positive and negative shocks need not have symmetric effects on conditional variance. This point is also consistent with Bomfim[5], who finds that the volatility response to monetary-policy

news is asymmetric. Although OLS remains useful as a transparent benchmark, it cannot fully account for fat-tailed errors, volatility clustering, and event-driven shocks. The present thesis, therefore, compares a simple linear framework with a GARCH specification, allowing the sentiment signal to enter a model that is better suited to the statistical properties of the VIX.

There is also a strand of literature on model complexity, domain adaptation, and robustness. Gururangan et al.[8] show that continued pretraining on domain-specific text can improve the performance of language models. Hu et al.[10] show that Low-Rank Adaptation can preserve much of the performance of large models, while substantially reducing the number of trainable parameters. These studies suggest that further specialization and parameter-efficient fine-tuning may improve financial-language modeling. However, the setting of the present thesis differs from the typical classification environments in which these methods are often evaluated. The corpus of FOMC minutes contains 206 documents spanning 2000-2025, the release frequency is low and the model is not evaluated by its ability to classify text into sentiment categories. Instead, the sentiment score is used as an explanatory variable in a financial-market regression, where the dependent variable is the five-day change in the VIX. Under these conditions, greater linguistic specialization may improve domain fit in the language space without improving explanatory power in asset-pricing space.

The present thesis contributes to this literature by examining whether increasing model complexity improves the economic usefulness of sentiment signals when textual information is sparse and the market outcome is noisy. It compares a simple LM lexicon approach with more complex transformer-based approaches, including FinBERT and domain-adapted variants. These sentiment signals are evaluated in the monetary-policy setting developed by the central-bank communication literature, and through volatility models motivated by the financial econometrics literature. The central argument is that complexity has value at different levels. In the natural-language-processing layer, greater complexity can generate a richer sentiment signal, but this does not necessarily translate into superior out-of-sample performance. Excessive domain adaptation may even reduce usefulness by "normalizing" the very language that markets treat as informative. In the econometric layer, however, complexity is more clearly justified, as financial time series returns distributions often exhibit fat tails and volatility itself is time-varying (heteroskedastic), meaning periods of high volatility tend to cluster. The contribution of the thesis is therefore to show that the optimal modeling approach depends on matching model complexity to the frequency, noise, and structure of the information environment.

3 Data and Event Construction

3.1 FOMC Document Sample

The textual dataset used in the thesis consists of the FOMC (Federal Open Market Committee) minutes covering the period 2000 - 2025. The sample contains 206 FOMC documents, each of which is assigned a release date and counted as a separate observation for analysis.

FOMC minutes are released every 8 weeks, and they follow a structured institutional

format. Therefore, the minutes of the FOMC meetings are appropriate for this analysis as they are published at a relatively low frequency, are fairly institutionalized in format, and include information regarding the Federal Reserve’s view on economic conditions, financial risks, and the potential future path of monetary policy.

3.2 VIX

The second variable used in this analysis is the CBOE Volatility Index (VIX). In financial terms, the VIX is a real-time index that measures the market’s expectation of 30-day volatility for the S&P 500, derived from S&P 500 option prices. The VIX is the dependent variable against which the different sentiment signals are evaluated. Normally, the VIX is calculated in real time during market hours, but for this thesis, a five-day forward change specification is used. The five-day specification is chosen to allow the market to ”digest” the information included in the minutes while also limiting the risk that the ”market response” will be affected by unrelated events.

3.3 Temporal Alignment and the Trading-Day Problem

The main problem that this thesis addresses and that the construction experiences is a frequency mismatch between the financial data and the textual data. FOMC minutes are released on specific dates every 8 weeks, while the VIX is only available on trading days.

To address this issue, an ”as-of” temporal alignment procedure is used, essentially, each FOMC document is matched to the next open-market trading day when the VIX is calculated. This procedure preserves the causal ordering of the analysis, as the market can digest the minutes only after the documents are released. It also avoids any potential look-ahead bias by ensuring that the signal is not matched to a VIX observation before the release of the corresponding minutes.

3.4 Event Window Construction

After each FOMC document is matched to the next valid trading day, the five-day change in VIX is calculated:

$$\Delta VIX_{t-1,t+5} = VIX_{t+5} - VIX_{t-1} \quad (1)$$

where t denotes the matched day following the FOMC document release. This transformation converts the VIX time series into an event window. Each FOMC document is therefore associated with one subsequent volatility response.

3.5 Descriptive Evidence and Event Coverage

This section presents descriptive evidence on the main variable used in the analysis and on the construction of the event dataset. Table 1 reports the descriptive statistics for the event-window VIX change. The mean change is -0.47, while the median is -0.4. The standard deviation, variance, range, and interquartile range all indicate variation in market movements.

Figure 1 compares the distribution of VIX levels on FOMC document release days with the unconditional distribution of VIX levels across all market days. The two distributions overlap substantially. This indicates that FOMC releases may not have a significant effect on the VIX, making it a neutral testing ground for comparing different models.

Figure 2 presents the daily VIX series alongside the matched FOMC document release dates, indicating that the sample spans a range of volatility regimes over time.

3.6 Summary of the Dataset

The final dataset consists of the low-frequency FOMC textual data and the high-frequency market data. In total, the sample contains 206 FOMC documents from 2000 to 2025, and each of them is matched to the next valid trading day. In the following analysis, the dependent variable is the 5-day change in VIX following each FOMC release day.

4 Methodology

This section explains the methodological framework used to transform FOMC minutes into measurable sentiment signals and then evaluate whether those signals contain information about short-run movements in market uncertainty. The analysis is divided into two main stages. The first is the Natural Language Processing (NLP) stage, where the textual data is converted into sentiment measures using both simple and more complex approaches. These include the Loughran-McDonald (LM) financial lexicon, pre-trained FinBERT, domain adaptation through Masked Language Modeling (MLM) and a parameter-efficient Low-Rank (LoRA) specification. The second is the econometric stage, where the generated sentiment signals are tested against the five-day change in the VIX. This is done first through a linear OLS benchmark and then through a GARCH framework, which is deemed more appropriate for volatility data.

4.1 Natural Language Processing Stage

This section describes the first stage of the methodological analysis, namely the NLP analysis.

4.1.1 Loughran–McDonald Approach

The baseline sentiment signal uses the Loughran–McDonald (LM) financial dictionary, which was introduced because general-purpose sentiment lexicons misclassify many words that are common in finance. In this thesis, the LM model is estimated in three versions (V). Version V1 entails minimal textual cleaning. Version V2 uses an NLTK-based pre-processing list augmented through Lasso selection.¹ Version V3 uses a more aggressive Elastic Net cross-validated filter.² In order to generate a sentiment signal, the thesis uses a bag-of-words scoring procedure, where each document is tokenized, and the number of

¹In NLP, Lasso (Least Absolute Shrinkage and Selection Operator) is an embedded feature selection technique used primarily for text classification and sentiment analysis. It removes irrelevant words and/or n-grams to prevent overfitting in high-dimensional text datasets.

²Elastic Net combines Lasso penalties for removing irrelevant features with penalties that handle multicollinearity without overfitting.

positive and negative words is counted according to the LM classification. The sentiment score is then constructed as the difference between positive and negative terms, scaled by the total number of relevant words in the document.

Formally, the LM-based sentiment score can be written as:

$$S_t^{LM} = \frac{Positive_t - Negative_t}{Total\ Words_t} \quad (2)$$

where $Positive_t$ and $Negative_t$ denote the number of positive and negative words in document t , and $Total\ Words_t$ refers to the total number of words after the corresponding preprocessing step.

4.1.2 FinBERT

The second sentiment signal is the contextual baseline built from a pre-trained FinBERT model. FinBERT adapts the BERT architecture to financial language. BERT itself was designed to learn bidirectional contextual representations rather than bag-of-words counts. In the thesis, the FinBERT score is defined as the difference between the model’s positive and negative softmax probabilities. A document can therefore be mostly neutral, mildly positive, or strongly negative depending on the relative distance between the positive and negative probabilities.

The FinBERT-based sentiment score can be written as:

$$S_{FinBERT} = P(pos) - P(neg) \quad (3)$$

where $S_{FinBERT}$ is the document-level sentiment score generated by the FinBERT model, $P(pos)$ is the softmax probability that the document expresses positive sentiment, and $P(neg)$ is the softmax probability that the document expresses negative sentiment.

4.1.3 FinBERT with Low-Rank (LoRA) Adaptation

The third sentiment signal uses LoRA adaptation. LoRA is a highly efficient technique used to adapt large AI models, like FinBERT, to specific tasks, such as financial sentiment classification. Instead of updating billions of parameters across FinBERT’s entire network, which entails a computational cost and the risk of model overfitting to the FOMC corpus, LoRA freezes the original weights. This means that it locks the model’s existing mathematical parameters into a permanent, read-only state, so they cannot be altered. LoRA then introduces a set of small, lightweight auxiliary matrices that remain unfrozen. These matrices act like a targeted mathematical filter, capturing the specific nuances of financial sentiment, while leaving FinBERT’s massive core knowledge untouched. During training, the computer only updates these small, adaptable matrices to teach them how to accurately classify text as positive, negative, or neutral. Once training is complete, these lightweight adjustments are mathematically blended back into the main FinBERT network. This creates a customized sentiment classifier that runs at peak performance without substantial delays.

This makes the adaptation process substantially lighter. In this thesis, the adapters are placed in the self-attention mechanism, specifically in the query and value projections,

since these components determine how the model assigns contextual importance across words. The adapted forward pass is written as:

$$W_0x + \frac{\alpha}{r}(BA)x \quad (4)$$

where W_0 represents the frozen pre-trained weights, x is the input representation, A and B are the trainable low-rank matrices, r is the chosen rank, and α is the scaling factor. Further details can be seen in Table 4

4.1.4 Masked Language Modeling (MLM)

The fourth sentiment signal is obtained through domain adaptation via MLM. Here, the thesis extends the base transformer by continuing pre-training on the FOMC corpus before sentiment scoring. The underlying idea comes from the adaptive pre-training literature. Specifically, models trained on broad corpora can often be improved by a second stage of in-domain pre-training. Operationally, the thesis follows the standard BERT-style MLM strategy, randomly masking 15% of tokens and training the model. Perplexity is tracked as a diagnostic measure of domain fit. Perplexity is a standard metric used to measure how well a language model predicts a sample of text, here the FOMC minutes. Lower perplexity means that the model better understands the vocabulary, tone, and context of the Fed, making fewer mistakes in sentiment analysis.

BERT does not read text in the same way as a traditional left-to-right language model. A traditional language model predicts the next word by looking only at the words that came before it. BERT, however is trained differently since it learns by using surrounding context.

For this reason the perplexity reported for BERT should not be interpreted as standard left-to-right perplexity. Instead it is used as a diagnostic tool based on the masked language modeling loss. Essentially, it helps show whether the continued pre-training helped BERT adapt to the financial domain.

Conventional left-to-right perplexity is defined as:

$$PPL(X) = \exp \left(-\frac{1}{N} \sum_{i=1}^N \log P(w_i|w_{<i}) \right) \quad (5)$$

where $PPL(X)$ is the perplexity of the document sequence X , N is the number of tokens, and $P(w_i|w_{<i})$ is the probability assigned by the model to token w_i , given the preceding context. After this continued pre-training step, sentiment is extracted using the same probability-difference logic as in the base FinBERT specification.

4.2 Econometric Stage

The econometric stage begins by converting the text output into an event-time dataset, as discussed in Section 3.

4.2.1 Ordinary Least Squares (OLS)

OLS is used as the baseline. The OLS specification is defined as:

$$\Delta VIX_{t-1,t+5} = \beta_0 + \beta_1 S_{i,t} + \epsilon_t \quad (6)$$

where $S_{i,t}$ is the sentiment signal generated by one of the NLP methods, β_0 is the intercept term, β_1 measures the correlation between the sentiment score and the five-day change in the VIX, and ϵ_t is the error term.

4.2.2 GARCH

The second econometric specification is the GARCH model, which is supported by the diagnostic results of the OLS baseline. Although the OLS regressions show that some sentiment signals are statistically associated with the five-day change in the VIX, the residual diagnostics indicate that the linear model does not fully capture the structure of the data. Across all sentiment specifications, the kurtosis values are high, ranging from 6.538 to 7.124, while the Jarque-Bera probabilities are below 0.001. This indicates that the residuals are fat-tailed and strongly non-normal. More importantly, the ARCH-LM test is statistically significant for all four sentiment signals, with p-values ranging from 0.0091 to 0.0204. This shows that conditional heteroskedasticity remains present after the OLS estimation. Therefore, the GARCH component is introduced to model the conditional variance directly. For further details on the diagnostic results see Table 7.

The conditional mean equation is specified as:

$$\Delta VIX_{t-1,t+5} = \mu + \delta S_{i,t} + \epsilon_t \quad (7)$$

where $\Delta VIX_{t-1,t+5}$ is the five-day change in the VIX following the release of the FOMC minutes, μ is the constant term, $S_{i,t}$ is the sentiment signal generated by NLP methods, and δ measures the relationship between the sentiment signal and the five-day change in the VIX. The conditional variance equation is specified as:

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \quad (8)$$

where σ_t^2 is the conditional variance of the error term at time t . The parameter ω is the long-run baseline variance. The parameter α is the ARCH coefficient that measures the effect of recent shocks, captured by the lagged squared residual ϵ_{t-1}^2 . The parameter β is the GARCH coefficient that captures the persistence of past volatility through the lagged conditional variance, σ_{t-1}^2 .

5 Results

This section presents the results from the NLP and econometric stages. The NLP results suggest that contextual models extract richer sentiment signals in comparison to the LM lexicon approach, as they move beyond simple word counts and capture a wider range of tone and context. However, this advantage is limited, especially when the model becomes too closely adapted to the FOMC corpus, where the signals weaken rather than improve, suggesting that domain specificity may remove parts that matter for market reactions. The econometric stage provides a clearer distinction. The OLS is useful as a first benchmark because it indicates whether sentiment is linearly related to the 5-day change in VIX. Still, the residual diagnostics indicate that the linear model does not fully capture the volatility of the data (see Table 7).

5.1 Natural Language Processing Results

The first comparison begins with the LM specifications. Table 2 and Figure 3 show the effect of tightening the preprocessing step from version V1 to version V3. Moving from V1 to V3 increases the raw standard deviation of the sentiment measure from 0.0074 to 0.0132, while the minimum value falls from -0.0357 to -0.0567. This implies that the Lasso and Elastic Net preprocessing filters help strip some of the verbal noise from the FOMC minutes, making the LM score more responsive to shifts in tone. Even so, the LM sentiment score remains around a negative average and very infrequently enters into positive territory. This suggests that the LM lexicon, though a useful benchmark, is unable to distinguish between Fed language formality and shifts to more pessimistic sentiment.

The comparison with the transformer baseline makes this point more clearly. Together Table 3 and Figure 4 are used to show whether the transformer-based approach provides additional variation and historical responsiveness beyond the dictionary-based LM measure. They show that the base FinBERT specification produces a more differentiated signal than the lexicon. Relative to LM-V3, the mean becomes more negative, while the standard deviation increases from 0.013 to 0.224 and its range expands from 0.073 to 1.410. The FinBERT sentiment series also moves into positive territory in periods such as 2003–2006 expansion, and it falls sharply during the 2008 crisis. By contrast, the LM series remains much closer to a persistently negative baseline. This suggests that the contextual model is better at separating institutional policy language from text that conveys a more distinct shift in tone.

After establishing that FinBERT produces a richer sentiment measure, the next question is whether this gain requires the full computational burden of the base FinBERT transformer. This is where the LoRA specification comes in. The purpose of LoRA is to test whether a more efficient version of FinBERT can preserve the same information while reducing the resources required to run the model. Table 4 and Table 5 summarize this trade-off while Figure 5 visualizes how inference time changes with document length across the LM lexicon, base FinBERT, and LoRA-FinBERT specifications. Together, these results show that the LoRA specification updates only a very small fraction of the total parameter space and reduces memory usage from 426.0 MB to 70.1 MB, although inference becomes slower, at about 6.9 seconds per document compared to 4.9 seconds for the base model. What is more important is that the parameter-efficient version does not produce stronger regression results than the base FinBERT. In the econometric tables that follow, the LoRA coefficients, fit measures, and diagnostics are identical to the base model. This means that LoRA is useful as a computational device because it preserves the FinBERT signal at lower memory cost without altering the main findings in any way.

The next step in complexity relates to the adaptation of FinBERT’s MLM approach. Figure 6 shows that perplexity declines sharply during continued adaptation and pre-training, indicating that the model does learn the statistical structure of the FOMC corpus. However, Figure 7 and Table 6 show that this extra adaptation is accompanied by a collapse in the sentiment signal itself. The mean increases from -0.205 to -0.017. The raw standard deviation falls from 0.224 to 0.008, and the positive upper tail nearly disappears. This suggests that, once the model is trained and adapted on FOMC language, words associated with risk, caution, and uncertainty start looking commonplace rather than informative. As a result, the model becomes less sensitive to sentiment variations that may

carry market relevance.

Taken together, the NLP results support the case for complexity up to a limit. Base FinBERT extracts a richer sentiment signal than the LM versions, as one would expect from a model that uses context rather than isolated word counts. The next step in complexity, however, does not continue this improvement. In fact, the results show that greater complexity via MLM adaptation does not always correspond to greater explanatory power.

5.2 Econometric Results

This section presents the results from the econometric stage of the analysis, which examines if Fed sentiment contains information about market uncertainty.

5.2.1 OLS

The OLS framework is used as a first benchmark. Table 7 presents the results from equation (6), which regresses the five-day change in the VIX on each sentiment measure. The coefficient on the base FinBERT sentiment is 2.8 and statistically significant at the 5 percent level, with the LM also showing a positive but statistically insignificant coefficient. Although this result may seem counterintuitive, positive FOMC language can still increase market uncertainty if investors interpret Fed optimism as a signal that the economy is strong enough to justify tighter monetary policy. In particular, optimistic language about macroeconomic conditions may lead market participants to anticipate stronger growth, potential overheating, or greater inflationary pressure. This, in turn, can raise expectations of less future accommodation, higher interest rates, or a more restrictive policy stance. As investors hedge against the risk of future policy tightening, demand for downside protection may increase, raising implied volatility and therefore the VIX despite the apparently positive tone of the announcement. The same is true for the LoRA computational-saving specification. Although the existing literature does not estimate the same FinBERT-sentiment-to-VIX relationship, this interpretation is consistent with Hubert and Labondance [11], who show that central-bank tone conveys information about the future path of monetary policy. If investors interpret optimistic FOMC language as a signal of less future accommodation or a higher probability of policy tightening, then positive sentiment may increase rather than reduce uncertainty. This mechanism is also consistent with Bomfim [5], who shows that monetary-policy surprises increase stock-market volatility in the short run and that higher-than-expected target-rate decisions generate larger volatility responses than lower-than-expected decisions.

The exception to this pattern is the MLM adapted specification, where the sign reversal may indicate over-specialization during pre-training. Because the model is further trained on the domain of the FOMC language, it may have become more sensitive to technical and cautious policy wording, while reducing the weight placed on generally positive language. As a result the MLM sentiment score may no longer be able to capture the positive tone the same way FinBERT does. Instead positive or higher sentiment produced by the MLM may reflect language more normalized or policy stabilizing within the FOMC corpus, which would, in turn, be associated with a lower VIX response.

5.2.2 GARCH

The main limitation of the OLS results is presented by the residual diagnostics in Table 7. For all four sentiment signals, kurtosis ranges from 6.5 to 7.1, and the Jarque–Bera probability is below 0.001. This indicates that the residuals are fat-tailed and strongly non-normal. More importantly, the ARCH-LM test is significant in all cases, with p-values ranging from 0.0091 to 0.0204. That means conditional heteroskedasticity remains in the residuals after estimation. By contrast, the Ljung–Box p-values are not significant, thus serial correlation is not an issue. For this reason Table 7 should be treated as a benchmark rather than a final specification. Figure 8 reinforces this limitation visually: the fitted values track the direction of some smaller movements in the VIX, but they do not capture the magnitude of the larger volatility spikes. Taken together, the diagnostic tests support the move from OLS to GARCH: the OLS residuals exhibit significant ARCH effects, while the GARCH specification largely removes remaining conditional heteroskedasticity. This empirical improvement is consistent with Apergis [2], who also uses GARCH volatility modeling to show that FOMC minutes contain information relevant for U.S. asset-price volatility.

The analysis therefore turns to the GARCH(1,1) specification, presented in Table 8 and Figure 9. There, the ARCH-LM p-values increase sharply, ranging from 0.82 to 0.98. This suggests that the GARCH specification matches the volatility clustering that the baseline OLS missed. The Ljung-Box p-values for standardized residuals are not statistically significant, indicating that serial correlation is not driving the remaining errors. Residual non-normality is still present, as the Jarque-Bera probabilities remain below 0.001 and kurtosis is still high. This is not surprising for VIX-related data, where extreme movements are common.

The coefficient on the base FinBERT signal increases to 3.5, compared to 2.8 from Table 7, and is statistically significant at the 1 percent level. Again, the LoRA specification produces the same results as FinBERT. For FinBERT and LoRA, the AIC falls from 859.6 in OLS to 849.6 in GARCH. The MLM-adapted model again has the worst performance, with the lowest R^2 , the highest MAE and RMSE, and the second highest AIC and BIC among the four specifications.

Another important result concerns the volatility equation itself. The ARCH coefficient is positive and tightly estimated, ranging from 0.27 to 0.36 across specifications. By contrast, the GARCH persistence term is effectively zero in every model. This suggests that the volatility response captured in the event window is immediate rather than persistent.

Overall, FinBERT remains the strongest in-sample specification. It has the lowest AIC and BIC among the GARCH models, and its sentiment coefficient is statistically significant at the 1 percent level. This is consistent with earlier NLP evidence, which showed that the base contextual model produced a more differentiated sentiment measure than the LM lexicon. At the same time, the LM model performs surprisingly well once it is placed in the correct econometric structure, as it has the lowest MAE and RMSE among the GARCH specifications. This means that, although the LM approach is weaker in terms of contextual richness and in-sample explanatory power, it remains competitive in prediction error.

6 Robustness Checks

This section examines whether the results are sensitive to the treatment of standard errors and to extreme movements in the five-day change in the VIX.

6.1 Robust Standard Errors (HC1)

The first robustness check re-estimates the OLS specification using heteroskedasticity-robust standard errors. The results are shown in Table 9. The sentiment coefficient remains positive and statistically significant after HC1, with the LoRA specification having the same results as the baseline FinBERT. Again, this suggests that, with low frequency data, parameter fine-tuning preserves the information contained in the base specification. By contrast, the LM and MLM approaches, which were marginally statistically significant under conventional standard errors, become insignificant. This strengthens the interpretation that FinBERT signals contain the most reliable relationship with a linear econometric specification.

6.2 Winsorized VIX

The second robustness test addresses the possibility that the baseline results are driven by extreme VIX movements. To address this concern, the five-day VIX change is winsorized at the 1st and 99th percentiles, and both the OLS and GARCH specifications are re-estimated. The results are shown in Table 10. In the winsorized OLS specification, FinBERT remains statistically significant at the 5 percent level, while the MLM and LM specifications are weaker. This pattern is similar to the baseline OLS results and suggests that the FinBERT relationship is not explained only by extreme observations in the dependent variable.

In the GARCH modeling, FinBERT remains the strongest in-sample specification, with the lowest AIC and BIC among GARCH models and a sentiment coefficient that is statistically significant at the 1 percent level. The MLM and LM coefficients are also statistically significant, highlighting the importance for inference of modeling conditional heteroskedasticity. The lexicon produces the lowest MAE and RMSE among the winsorized GARCH specifications, indicating again that simple sentiment measures should not be dismissed in low-frequency environments.

6.3 EGARCH Specification

This section examines if the previous GARCH results are robust to adding asymmetric volatility responses. In this case, the standard GARCH(1,1) specification is modified to allow for different asymmetric effects of negative and positive shocks on conditional volatility, commonly known as the leverage effect, by using an EGARCH(1,1) specification. The extension could be significant as VIX could react in varying ways to contractionary or expansionary FOMC statements.³

³The leverage effect is the tendency for volatility to increase more sharply in response to negative returns (bad news) than to positive returns (good news) of the same magnitude.

The results in Table 11 show that allowing for asymmetry does not materially change the main conclusions. FinBERT and LoRA remain the strongest specifications in terms of in-sample fit, while the LM model remains competitive in forecasting error. The MLM-adapted model continues to perform weakest overall. Although the EGARCH specification provides marginal gains in some information criteria, the improvements are small and do not overturn the baseline ranking of the models as seen in Table 11. While the baseline GARCH specification suggests limited persistence, the EGARCH robustness check indicates that persistence may reappear once asymmetric volatility responses are allowed. Therefore, the interpretation of the volatility dynamics merits further study.

This supports the broader interpretation that econometric complexity is useful for the NLP and econometric methodologies used, when it corrects a clear feature of the data, such as conditional heteroskedasticity. However, additional complexity beyond the GARCH framework produces only limited incremental benefits in this low-frequency setting.

7 Conclusion

In this paper, FOMC minutes from 2000 to 2025 are used to determine whether a higher level of NLP model complexity would enhance the power of financial text to account for VIX movements in a low-frequency setting. The findings indicate that switching from a lexicon model to base FinBERT enhances the sentiment signal and strengthens the link with volatility, particularly under GARCH. The results indicate that the MLM-adapted FinBERT model performs worse than the baseline FinBERT model, possibly because excessive domain adaptation loses the linguistic variation important for market reactions. In contrast, the lexicon is quite competitive in spite of its simplicity, especially when it is combined with the GARCH specification. The findings also demonstrate that the selection of econometric models is important. Although informative as a benchmark, OLS is not well suited to volatility, and the GARCH framework provides more reliable inference and better statistical results.

Overall, the evidence indicates that the most complex model at each stage is not necessarily the optimal specification in this setting. Instead, the best-performing approach combines a generalized contextual NLP model with an econometric framework that captures volatility behavior. However, this conclusion should be interpreted cautiously since the model ranking is specific to short-horizon VIX responses to FOMC minutes during the post-2000 sample and should not be read as evidence that simpler or LM-based sentiment measures are generally superior to BERT-based models across other financial or macroeconomic prediction tasks.

Future research could further compare the use of alternative event windows or the application of the same framework to higher frequency central-bank communication. Such extensions would help us determine whether the limited predictive power observed here is due to the specific nature of the FOMC minutes or a general limitation of complex NLP approaches for low-frequency financial text.

References

- [1] M. Apel and M. B. Grimaldi. How informative are central bank minutes? *Review of Economics*, 2014.
- [2] N. Apergis. The role of fomc minutes for us asset prices before and after the 2008 crisis: Evidence from garch volatility modeling. *The Quarterly Review of Economics and Finance*, 2015.
- [3] D. Araci. Finbert: Financial sentiment analysis with pre-trained language models. 2019.
- [4] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 1986.
- [5] A. N. Bomfim. Pre-announcement effects, news effects, and volatility: Monetary policy and the stock market. *Journal of Banking & Finance*, 2003.
- [6] R. F. Engle and V. K. Ng. Measuring and testing the impact of news on volatility. *The Journal of Finance*, 1993.
- [7] R. S. Gürkaynak, B. Sack, and E. T. Swanson. Do actions speak louder than words? the response of asset prices to monetary policy actions and statements. *International Journal of Central Banking*, 2005.
- [8] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don’t stop pretraining: Adapt language models to domains and tasks.
- [9] S. Hansen, M. McMahon, and A. Prat. Transparency and deliberation within the fomc: A computational linguistics approach. *The Quarterly Journal of Economics*, 2018.
- [10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. 2021.
- [11] P. Hubert and F. Labondance. The signaling effects of central bank tone. *European Economic Review*, 2021.
- [12] T. Loughran and B. McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 2011.
- [13] D. O. Lucca and E. Moench. The pre-fomc announcement drift. *The Journal of Finance*, 2015.
- [14] D. O. Lucca and F. Trebbi. Measuring central bank communication: An automated approach with application to fomc statements. Technical report, National Bureau of Economic Research, 2009.
- [15] R. C. Tadle. Fomc minutes sentiments and their impact on financial markets. *Journal of Economics and Business*, 2022.
- [16] P. C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 2007.

8 Appendix

Table 1: Descriptive Statistics for VIX

Mean	Median	SD	Variance	Range	IQR	Skewness	Kurtosis
-0.47	-0.4	3.34	11.18	28.09	3.0	-0.39	4.05

Table 1 shows descriptive statistics for the VIX over the period 2000-2025. See discussion in Section 3.5

Table 2: Marginal Utility of Stopword Removal

Version	Signal Intensity (σ_{raw})	Mean (μ)	Minimum	Maximum
V1 (Base)	0.0074	-0.0106	-0.0357	0.0056
V2 (2k)	0.0123	-0.0145	-0.0518	0.0183
V3 (3.2k)	0.0132	-0.0162	-0.0567	0.0162

Table 2 shows the change in Signal intensity, mean, Minimum and the Maximum based on the stopwords included for Loughran-McDonald. See discussion in Section 5.1

Table 3: Comparison: LM vs Base FinBERT

Metric	Lexicon (V3)	FinBERT (Base)	Delta / Factor
Mean (μ)	-0.016	-0.205	12.8 times bigger
Volatility (σ_{raw})	0.013	0.224	17.2 times bigger
Range	0.073	1.410	19.3 times bigger
Minimum	-0.057	-0.803	14.1 times bigger
Maximum	+0.016	+0.607	37.9 times bigger

Table 3 shows the amplification of the statistical metrics when using the Base FinBERT model. See discussion Section 5.1

Table 4: LoRA Hyperparameter Configuration.

Parameter	Value
Rank (r)	8
Scaling Factor (α)	16
Target Modules	query, value (in Self-Attention)
Dropout	0.1
Trainable Parameters	$\approx 0.2\%$ of Total

Table 4 displays the Hyperparameters used in the LoRA Specification. See discussion in Section 4.1.3 and 5.1

Table 5: **Optimization Trade-offs: LM vs FinBERT Base vs LoRA.**

Model Configuration	Params (Active)	Time/Doc (s)	Memory (MB)	Trade-off
LM Lexicon	0	0.0008	32.7	Baseline
FinBERT (Base)	109,484,547	4.9099	426.0	High Memory
FinBERT + LoRA	Rank-8 Adapters	6.8937	70.1	Slower / Lighter

Table 5 reports the computational requirements of LM lexicon, base FinBERT, and LoRA + FinBERT specifications. See discussion Section 5.1

Table 6: **Base vs Domain-Specification.**

Metric	FinBERT (Base)	FinBERT (MLM-Adapted)	Delta
Mean (μ)	-0.205	-0.017	91% Reduction
Volatility (σ_{raw})	0.224	0.008	96% Collapse
Minimum	-0.803	-0.058	
Maximum	+0.607	+0.004	

Table 6 shows the effect of domain-specific adaptation on the sentiment signal. See discussion in Section 5.1

Table 7: Baseline OLS Results: Sentiment Signals vs 5-Day Δ VIX

Metric	FinBERT (S_{Base})	LoRA (S_{LoRA})	MLM (S_{MLM})	LM (S_{LM})
R^2	0.040	0.040	0.018	0.018
Constant	0.2171	0.2171	-1.2484**	0.1510
Coefficient (β_1)	2.8158**	2.8158**	-53.2692*	34.8734*
Kurtosis	6.538	6.538	7.124	7.053
JB Prob.	< 0.001	< 0.001	< 0.001	< 0.001
MAE	2.3788	2.3788	2.4956	2.3697
RMSE	3.2847	3.2847	3.3787	3.2497
AIC	859.5688	859.5688	863.2208	863.3511
BIC	865.7686	865.7686	869.4205	869.5508
ARCH-LM test	p-value=0.0091	p-value=0.0091	p-value=0.0204	p-value=0.0141
Ljung-Box test	p-value=0.3809	p-value=0.3809	p-value=0.4548	p-value=0.4634

Notes: dependent variable is VIX_Change_5D . Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7 displays the metrics produced by the OLS regression. See discussion in Section 5.2.1

Table 8: GARCH(1,1) Results: Sentiment Signals vs 5-Day Δ VIX

Metric	FinBERT (S_{Base})	LoRA (S_{LoRA})	MLM (S_{MLM})	LM (S_{LM})
Constant	0.2940	0.2940	-1.8777	0.3334
Sentiment Coefficient (δ)	3.4852***	3.4852***	-92.0667**	48.5350**
ω	7.6410*	7.6410*	7.3361	8.1543
ARCH (α)	0.2991	0.2991	0.3555	0.2621
GARCH (β_{vol})	3.58×10^{-16}	3.58×10^{-16}	0.0000	0.0000
R^2	0.0374	0.0374	0.0086	0.0148
MAE	2.3689	2.3689	2.5639	2.3598
RMSE	3.2851	3.2851	3.4762	3.2605
AIC	849.5934	849.5934	849.6988	854.7058
BIC	865.0928	865.0928	865.1981	870.2052
JB Prob.	< 0.001	< 0.001	< 0.001	< 0.001
Kurtosis	7.1401	7.1401	6.6997	7.9819
ARCH-LM statistic	4.0980	4.0980	5.9190	3.1731
ARCH-LM p-value	0.9428	0.9428	0.8220	0.9771
Ljung-Box p-value, standardized residuals	0.5935	0.5935	0.4647	0.4020

Notes: dependent variable is VIX_Change_5D . Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Stars are applied to coefficient estimates based on their corresponding p-values.

Table 8 displays the metrics produced by the GARCH(1,1). See discussion in Section 5.2.2

Table 9: OLS Results with HC1 Robust Standard Errors: Sentiment Signals vs 5-Day Δ VIX

Metric	FinBERT (S_{Base})	LoRA (S_{LoRA})	MLM (S_{MLM})	LM (S_{LM})
Constant	0.2171	0.2171	-1.2484*	0.1510
Sentiment Coefficient (β_1)	2.8158**	2.8158**	-53.2692	34.8734
Robust SE	1.2032	1.2032	33.0509	22.3898
R^2	0.040	0.040	0.0183	0.0176
MAE	2.3788	2.3788	2.4956	2.3697
RMSE	3.2847	3.2847	3.3787	3.2497
AIC	859.5688	859.5688	863.2208	863.3511
BIC	865.7686	865.7686	869.4205	869.5508

Notes: dependent variable is VIX . Robust standard errors use HC1 correction.

Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Stars are based on HC1 robust p-values.

Table 9 reports the OLS results using heteroskedasticity-robust HC1 standard errors. See discussion in Section 6.1

Table 10: Winsorized OLS and GARCH(1,1) Results: Sentiment Signals vs 5-Day Δ VIX

Metric	FinBERT (S_{Base})	LoRA (S_{LoRA})	MLM (S_{MLM})	LM (S_{LM})
Panel A: Winsorized OLS				
Constant	0.1460	0.1460	-1.1582**	0.0893
Sentiment Coefficient (β_1)	2.4825**	2.4825**	-47.6998*	30.8568
R^2	0.0352	0.0352	0.0167	0.0156
MAE	2.3325	2.3325	2.4377	2.3228
RMSE	3.1316	3.1316	3.2269	3.1005
AIC	839.9632	839.9632	843.0815	843.2623
BIC	846.1629	846.1629	849.2812	849.4620
Panel B: Winsorized GARCH(1,1)				
Constant	0.1326	0.1326	-1.6164***	0.1996
Sentiment Coefficient (β_1)	2.9669***	2.9669***	-73.3898**	42.1986**
ω	6.3121**	6.3121**	6.3630	6.6463
ARCH (α)	0.3384**	0.3384**	0.3508*	0.3155
GARCH (β_{vol})	0.0000	0.0000	6.45×10^{-13}	2.10×10^{-15}
R^2	0.0326	0.0326	0.0117	0.0131
MAE	2.3132	2.3132	2.4718	2.3055
RMSE	3.1173	3.1173	3.2806	3.1011
AIC	826.0493	826.0493	828.4761	830.5246
BIC	841.5486	841.5486	843.9754	846.0240

Notes: dependent variable is VIX_Change_5D . Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Stars are applied to coefficient estimates based on their corresponding p-values.

Table 10 reports the winsorized OLS and GARCH(1,1) results. See discussion in section 6.2

Table 11: Asymmetric GARCH Robustness Results: Sentiment Signals vs 5-Day Δ VIX

Metric	FinBERT (S_{Base})	LoRA (S_{LoRA})	MLM (S_{MLM})	LM (S_{LM})
Constant	0.2058	0.2058	-1.7256	0.2222
Sentiment Coefficient (δ)	3.2410***	3.2410***	-83.5782**	43.0803**
ω	0.0866*	0.0866*	2.3425	0.0758
ARCH (α)	0.0564	0.0564	0.5093*	0.0745
Asymmetry (γ)	0.1793***	0.1793***	-0.0877	0.1336
GARCH (β_{vol})	0.9674***	0.9674***	0.0000	0.9757***
R^2	0.0382	0.0382	0.0123	0.0164
MAE	2.3638	2.3638	2.5529	2.3553
RMSE	3.2757	3.2757	3.4571	3.2522
AIC	848.1663	848.1663	853.9303	857.0556
BIC	866.7655	866.7655	872.5295	875.6548
JB Prob.	< 0.001	< 0.001	< 0.001	< 0.001
Kurtosis	6.3254	6.3254	6.7217	7.1683
ARCH-LM statistic	9.6236	9.6236	5.8370	10.5753
ARCH-LM p-value	0.4741	0.4741	0.8288	0.3916
Ljung-Box p-value, standardized residuals	0.8694	0.8694	0.4733	0.6988

Notes: dependent variable is VIX_Change_5D . Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Stars are applied to coefficient estimates based on their corresponding p-values.

Table 11 reports the asymmetric GARCH robustness results. See discussion in Section 6.3

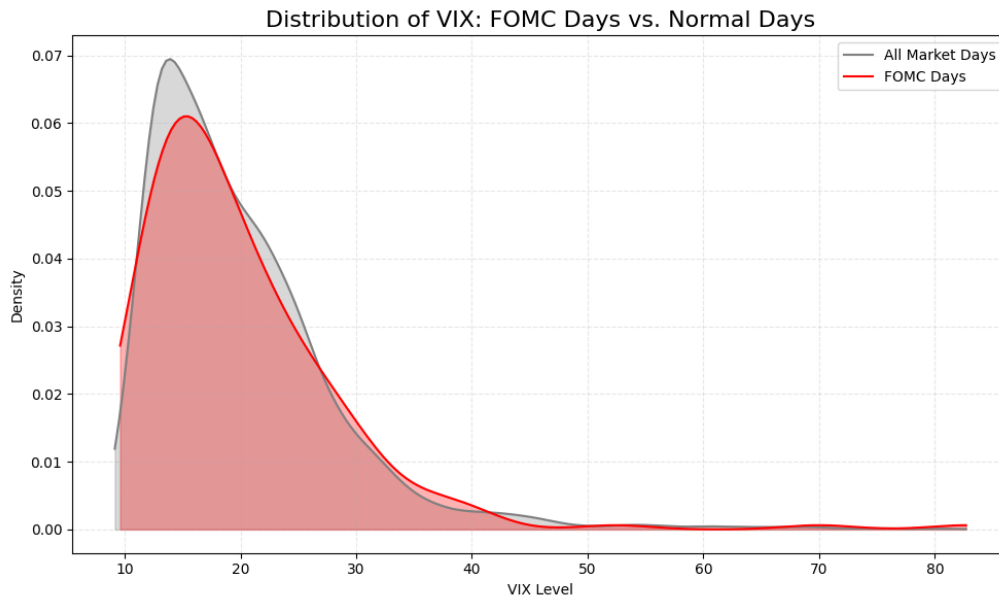


Figure 1 compares the distribution of VIX (solid line) levels on FOMC document release days (dotted points). See discussion in Section 3.5

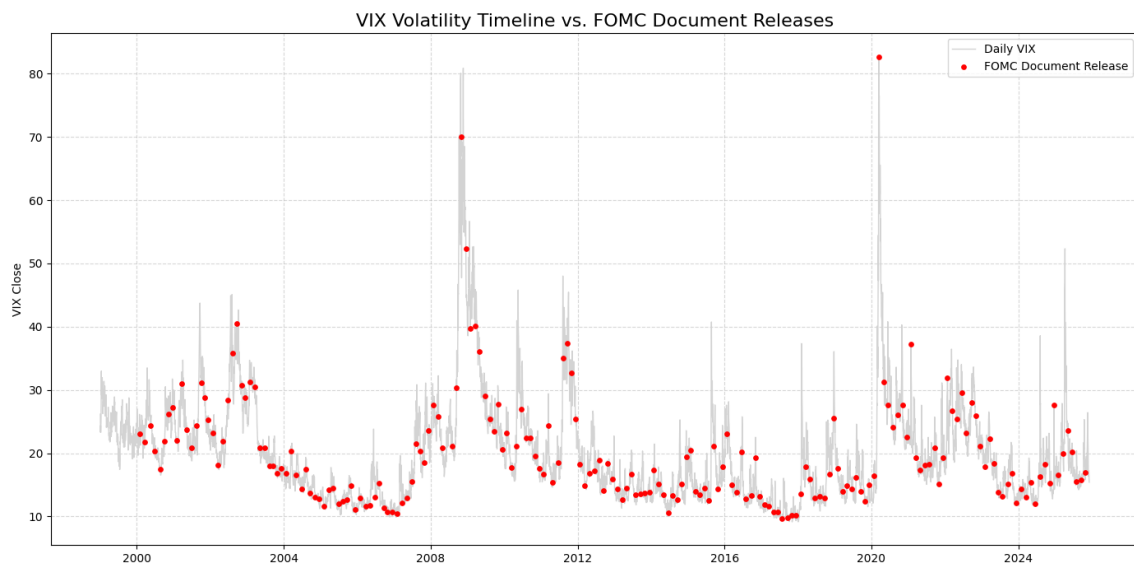


Figure 2 displays the Daily VIX and Matched FOMC Document Release Dates. See discussion in Section 3.5

Comparison of LM Sentiment Timelines

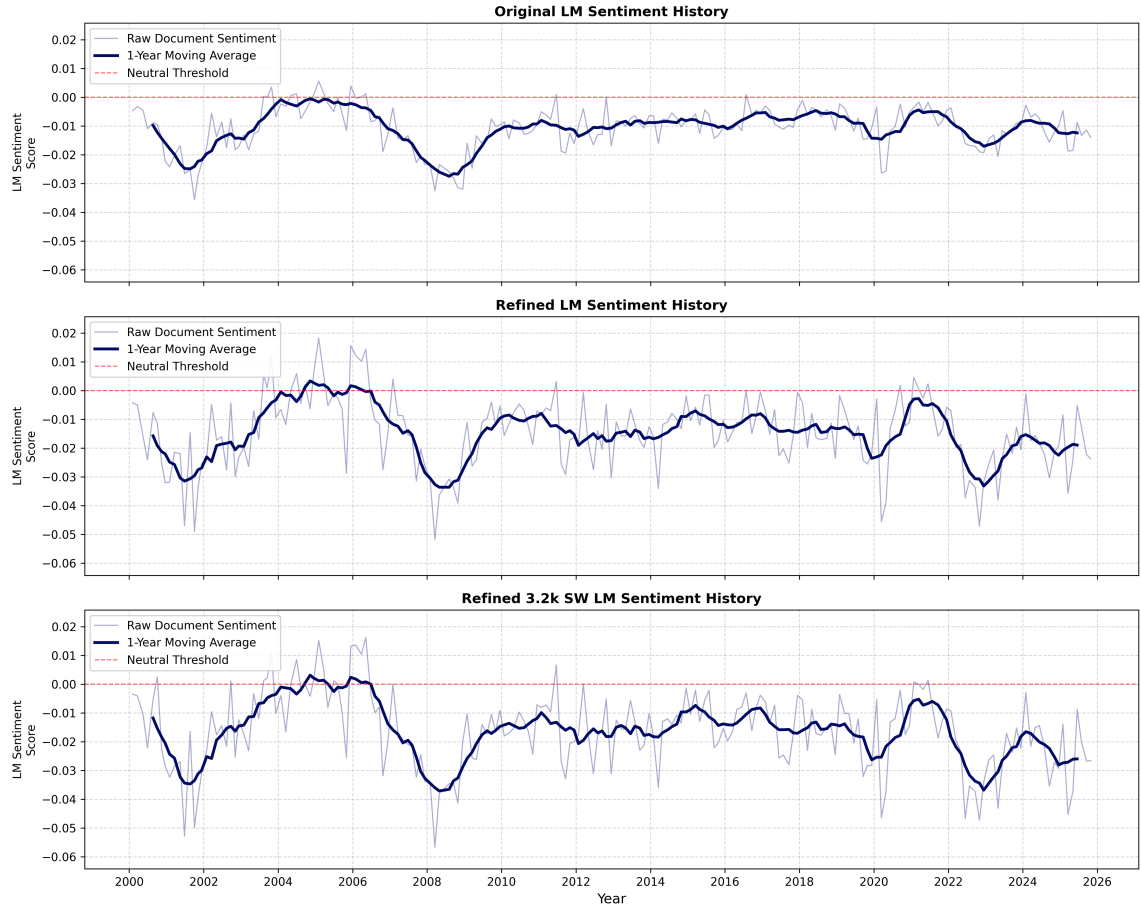


Figure 3 displays Loughran-McDonald sentiment scores from 2000 to 2025 under three preprocessing specifications or versions. The light blue line shows the raw sentiment score and the dark blue line shows the one-year moving average. See discussion in Section 5.1.

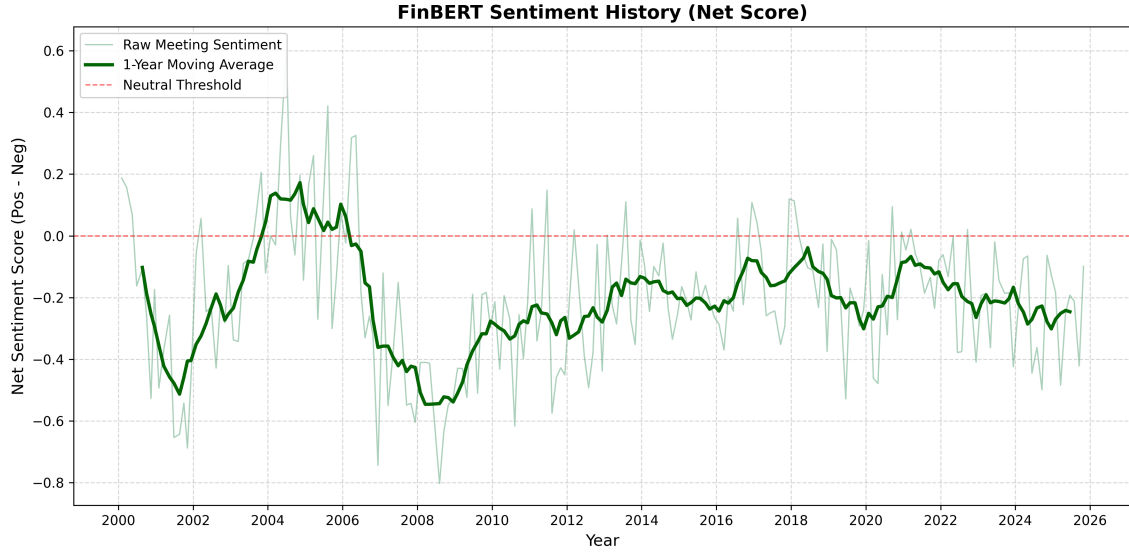


Figure 4 displays the sentiment scores when using the Base FinBERT specification from 2000 to 2025. The light green line shows the raw sentiment score and the dark green line shows the one-year moving average. See discussion in Section 5.1.

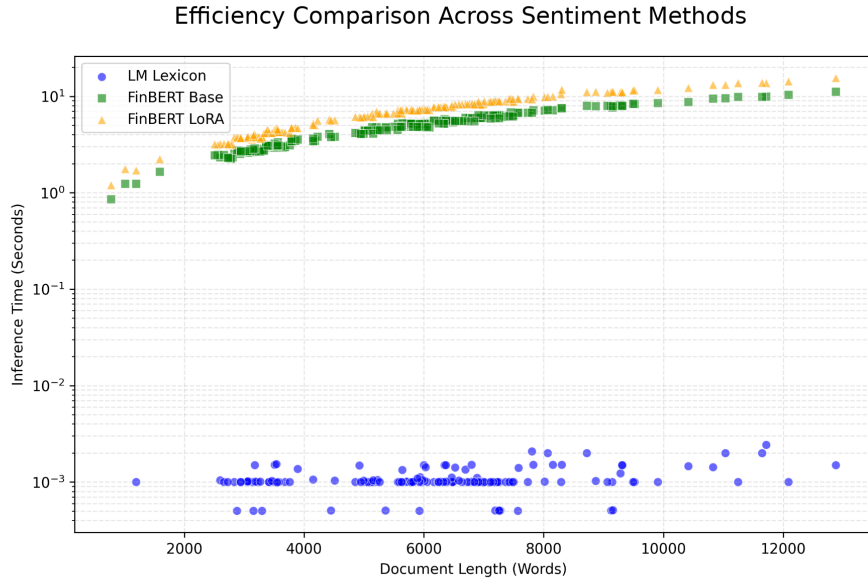


Figure 5 shows how processing time changes as document length increases for the LM dictionary method(blue dots), the Base FinBERT model(yellow triangles), and the LoRA fine-tuned FinBERT model(green squares). The vertical axis represents the time needed to process each document, while the horizontal axis represents document length. See discussion in Section 5.1

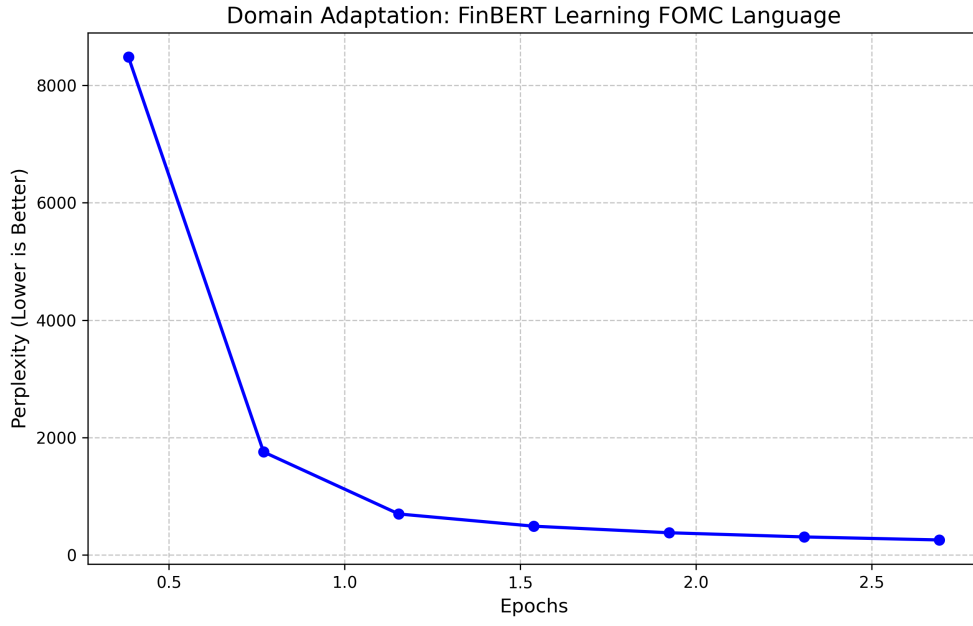


Figure 6 shows the evolution of perplexity during the Masked Language Modeling adaptation of FinBERT for the FOMC documents. See discussion in Section 5.1

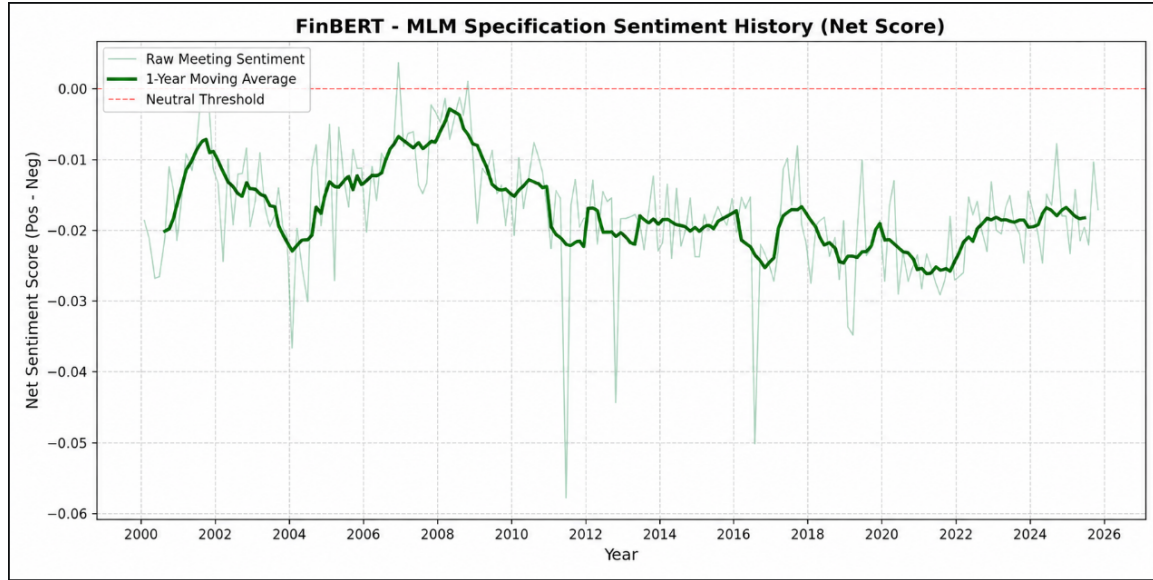


Figure 7 displays the sentiment scores when using domain-specific adaptation from 2000 to 2025. The light green line shows the raw sentiment score and the dark green line shows the one-year moving average. See discussion in Section 5.1

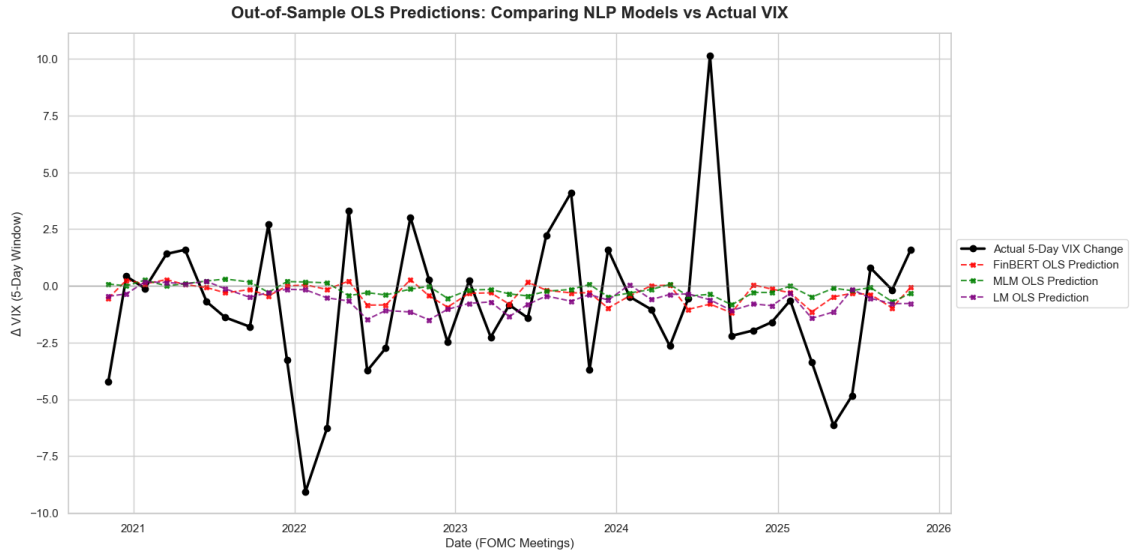


Figure 8 shows the out-of-sample OLS predictions using sentiment signals from the NLP models. The black line reports the actual five-day change in VIX, while the dashed lines show the predicted values from FinBERT, MLM, and LM. The LoRA specification is not shown because its predictions are identical to the FinBERT specification. See discussion in Section 5.2.1

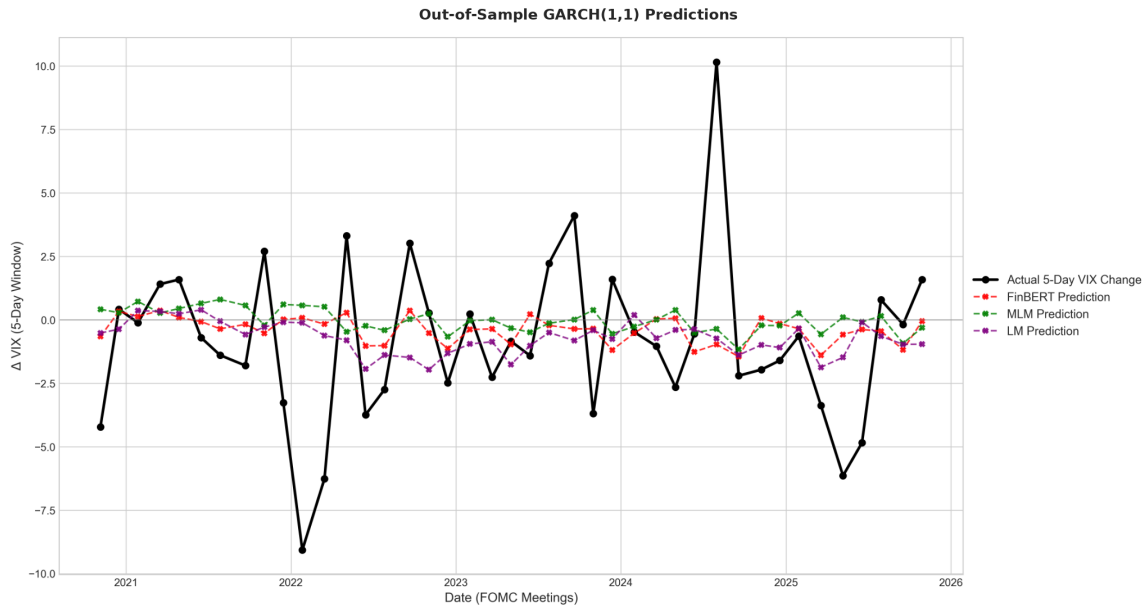


Figure 9 shows the out-of-sample GARCH predictions using sentiment signals produced from the 3 different NLP specifications. The black line represents the actual five-day change, with the colored dashed lines representing the predicted values from Base FinBERT, LM and MLM specifications. The LoRA specification is not shown because its predictions are identical to the FinBERT specification. See discussion in Section 5.2.2