Text Data Measured with Error: Empirical Strategies with an Application in S&P 500 Implied Volatility Forecasting

by

Nicolas van Hell

An essay submitted to the Department of Economics in partial fulfillment of the requirements for the degree of Master of Arts

Queen's University

Kingston, Ontario, Canada

August 2021

Abstract

This paper uses a parametric Heterogeneous Autoregressive [HAR] model augmented with central bank public speech sentiment to forecast S&P 500 implied volatility (CBOE VIX). Text sentiment computed by Natural Language Processing [NLP] algorithms may be measured with error due to estimation inaccuracy which leads to statistical bias and inconsistency in OLS estimation. As such, I propose two alternative methods to incorporate sentiment measures and contrast their performance through out-of-sample performance and the Diebold-Mariano [DM] test. Specifically, I examine both Instrumental Variable [IV] and Factor Analysis [FA] approaches to handle measurement error. With a sample of 5,449 trading days and 1,189 US Federal Reserve speeches, I find that (1) implementing financial and macroeconomic variables results in similar predictive ability compared to the base HAR, (2) integrating policymaker speech sentiment jointly with financial and macroeconomic variables significantly increases predictive ability over the HAR under the Diebold-Mariano [DM] test, (3) sentiment is more effective over short-term forecasts, becoming less effective as the forecast horizon expands, (4) IV and FA methods do not achieve significant outperformance against a single sentiment measure, and (5) the choice of test diagnostic is important when comparing forecast performance.

Acknowledgements

I would like to express my sincere gratitude to my advisor, Professor Steven Lehrer, for his invaluable insights, ideas, and feedback throughout this project.

Contents

1 Introduction

Consumer and investor sentiment are hypothesized to be important drivers of a variety of financial and macroeconomic phenomena. Long ago, Keynes (1936) proposed the idea of "animal spirits", where human psychology is a crucial determinant of financial decision making. Consequently, some forms of sentiment could be key variables in modeling financial and macroeconomic behaviour among economic agents. The usage of various forms of sentiment in empirical research is increasingly common largely due to rapid advances in Natural Language Processing [NLP] which have increased both accuracy in sentiment measurements and computational efficiency. However, testing sentiment-based theories is still challenged by limitations and methodological issues surrounding qualitative data, particularly when computing sentiment measurements from text data. This paper operates under the premise that sentiment measured through NLP methods is not error-free, and examines potential strategies to address measurement error in the context of modeling financial market volatility. In addition, this paper differs from previous research in that the sentiment of central bank policymakers is studied instead of more commonly examined types of sentiment, such as consumer or investor sentiment.

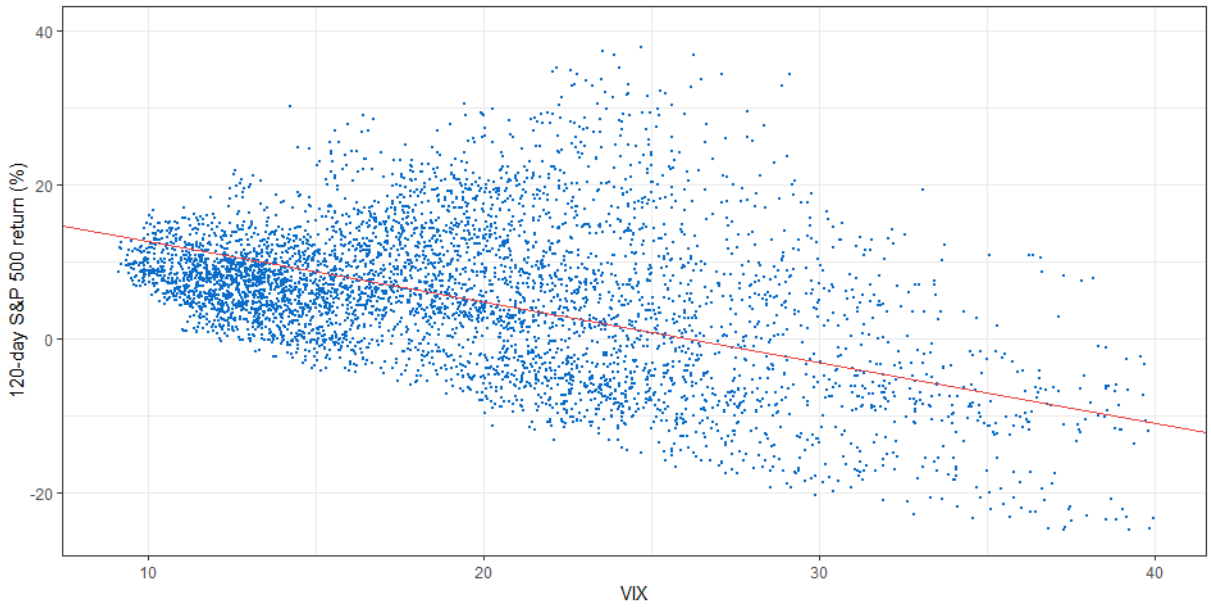
The remainder of this section is organized as follows. The motivation behind the choice of modeling volatility is introduced in subsection 1.1, the link between volatility and policymaker sentiment is expanded upon in subsection 1.2, and subsection 1.3 delves into some of the recent and relevant computer science literature related to sentiment analysis.

1.1 Implied Volatility

Price discovery is a central component of capital markets and crucial to the well-functioning of the financial sector. Risk, however, is widespread, and often shows up in the form of rapid changes in asset prices – particularly for equity-backed securities. At the extreme, a high degree of market volatility is strongly associated with more broad measures of macroeconomic challenges, often represented by substantial fear and uncertainty among economic agents (Chicago Board Options Exchange 2021). As such, it may be of particular interest to investors, institutions, and policymakers to better understand the drivers of volatility in asset prices.

There are many different methods to measure volatility in financial markets. This paper focuses on the CBOE Volatility Index (VIX), developed by the Chicago Board Options Exchange [CBOE], which measures the S&P 500 30-day market implied volatility by averaging S&P 500 put and call prices. Chicago Board Options Exchange (2021) highlights many advantages of the VIX. For instance, the VIX is not dependent on any particular option pricing model due to it being purely options-based. The index is also commonly considered by practitioners to loosely measure investor fear, some considering it a "fear gauge" of sorts. The VIX is therefore widely used to assess market conditions, particularly due to its implementability in investment portfolios. Fernandes et al. (2014) provides evidence and finds market returns are negatively correlated with the VIX, suggesting that high levels of stock market volatility may conditionally imply lower expected stock returns. To illustrate, Figure 1 demonstrates a clear negative univariate relationship between the VIX and 120-day S&P 500 returns using daily rolling periods between 1996-2018 from subsection 2.3.

Figure 1: Scatter Plot Between VIX and 120-day S&P 500 returns (1996-2018)



Because of the negative relationship between implied volatility and market returns, some trading strategies have emerged based on the VIX. For example, Clements and Fuller (2012) find that combining long positions in equities and implied volatility increases risk-adjusted performance. This is due to market turmoil simultaneously decreasing equity prices while also increasing implied volatility, suggesting the VIX potentially being an effective hedge against bear markets.

A large interest in forecasting implied volatility has emerged among practitioners since the VIX's inception, and various modeling approaches have been adopted by researchers. For example, Jiang and Lazar (2020) use a Generalized Autoregressive Conditional Heteroskedasticity [GARCH] model and finds strong model performance for both in-sample and out-of-sample VIX forecasts. Fernandes et al. (2014) test variants of the Heterogeneous Autoregressive [HAR] model and find it to be effective due to the VIX's highly persistent nature. Some approaches to VIX forecasting involve integrating various forms of sentiment measured by NLP algorithms.

Lehrer et al. (2021) use deep learning methods to incorporate twitter sentiment as a proxy for consumer confidence in a HAR-based model. The authors find that sentiment significantly increases forecast accuracy, highlighting the potential benefits of sentiment inclusion. Furthermore, Shvimer et al. (2021) model the VIX using sentiment extracted from economic articles. They use a Long Short-Term Memory [LSTM] text model combined with an Autoregressive Integrated Moving Average [ARIMA] model and find a 5% improvement in out-of-sample forecasting accuracy relative to plain ARIMA.

Although numerous studies, including Lehrer et al. (2021) and Shvimer et al. (2021), find that sentiment can improve VIX forecasts, the type of sentiment is important. Wang et al. (2005) examine whether investor sentiment is a useful variable for forecasting volatility and find that both market returns and volatility cause investor sentiment – not the other way around. This indicates that it is more useful to directly include stock returns and lagged volatility in a forecasting model than to use investor sentiment, a noisy proxy. Importantly, this result suggests that it is crucial to think carefully about the causality of measured sentiment and the dependent variable being modeled.

This paper extends current research in two ways. First, sentiment from central bank policymaker speeches is extracted to test whether its implementation in a HAR model improves implied volatility forecasts. Second, most current sentiment analysis research is based on assuming, at least implicitly, that sentiment is accurately measured. However, sentiment is estimated and thus likely contains error. To illustrate, NLP algorithms will compute differing sentiment measures over the same underlying text despite identical training data, implying that their measurements of true sentiment cannot all be simultaneously error-free. Many estimators become biased under measurement error. For example, OLS estimation in a HAR-type

model becomes biased towards zero. Therefore, I propose both instrumental variable and factor analysis methods to address measurement error and test these against uncorrected measures of sentiment in the context of forecasting the VIX.

1.2 Central Bank Policymaker Sentiment

The motivation for using central bank sentiment in the context of modeling implied volatility is worth highlighting. Stock prices are hypothesized to be a reflection of expected future discounted cash flow, and influenced by a myriad of determinants. One important determinant of market expectations surrounding future risky cash flows is monetary policy. Consequently, many US investors pay special attention to the Federal Reserve and its communications to the public. For instance, Vähämaa and Äijö (2010) find uncertainty in equity markets to be heavily influenced by the Federal Reserve, especially through FOMC meetings. Canadian evidence from Young Chang and Feunou (2014) suggests that both realized and implied volatility decrease, reflecting lower future uncertainty, when the Bank of Canada policy rates are announced. In addition, the impact of central bank policy on equity markets is substantial in magnitude. Bernanke and Kuttner (2005) estimate a 25 basis-point cut in the Fed's policy rate is associated with a 100 basis-point increase in stock prices. The authors find that monetary policy primarily affects stock prices through its influence on expected future dividends and expected future excess stock returns, as opposed to being influenced directly by real interest rates. For example, tight money can increase the risk surrounding stocks by weakening firm balance sheets and increasing risk-aversion, both of which raise the equity risk premium. However, they also find that the majority of the variation in stock prices is due to non-monetary factors. This last result might point to limitations of using sentiment derived from central bank policymakers in VIX modeling.

The linkage between variation in central bank policy and uncertainty in financial markets has the potential to be studied. Specifically, the VIX is considered to be a proxy for fear and uncertainty. Therefore, communications from the central bank towards the public may be useful in forecasting indicators of fear and uncertainty, such as the VIX. That said, the literature is unclear on the usefulness of policymaker sentiment in VIX forecasting largely due to limited existing research in this area.

1.3 NLP and Sentiment Analysis

Text data is increasingly common in economic research. Gentzkow et al. (2017) provide a detailed introduction to its uses in economics while outlining relevant statistical methods for applied work. The richness of text data is highlighted by the authors; a large quantity of human interactions are taking place digitally which offers great potential to better understand social and economic activities. Moreover, due to rapid advances in the field of NLP, the ability to capture information from text data has increased dramatically. Most economic research, however, currently focuses on implementing a single NLP model to make quantitative measurements from the data. In the context of sentiment analysis, this suggests that sentiment measures may contain error. This paper instead contrasts four state-of-the-art text models and applies both IV and FA methods to address possible measurement error. Recent computer science literature related to these four text models is further detailed in the remaining of this section.

One important semi-recent NLP breakthrough is the Transformers architecture developed by Vaswani et al. (2017) at Google AI. Transformers introduce substantial improvements in the way text is modeled, particularly with the introduction of self-attention. In short, self-attention

computes representations of a given token[1] sequence by relating the positions of every token with each other. As a result, the neural network is able to focus on the most important relations between embedded tokens to better capture linguistic context. In addition, Transformers greatly reduce computation time by allowing parallelization[2]. This architecture forms the basis of many of the newest and most performant NLP text models.

Based on the Transformer architecture, the Bidirectional Encoder Representations from Transformers [BERT] language model introduced by Devlin et al. (2019) marked an important turning point in NLP, considered by many a new era of text modeling. It achieved records in 11 natural processing tasks, including strong performance in sentiment analysis which is of particular relevance to this paper. BERT uses Masked Language Modeling [MLM] which masks token inputs at random. Its deep bidirectional[3] Transformer is pre-trained by predicting each masked input using text data from both English Wikipedia (2.5 billion words) and the BooksCorpus by Zhu et al. (2015) (800 million words). A notable improvement over previous models is BERT's bidirectional architecture in place of a unidirectional architecture which enables it to capture context from both the left and right of each token, better modeling the context surrounding the token. Soon after BERT was published by Google AI's team, Facebook AI published RoBERTa, developed by Liu et al. (2019). The authors found BERT to be undertrained and the choice of hyperparameters[4] have large impacts of performance. RoBERTa improved the design of BERT through better training data and hyperparameter optimization which led to a new state-of-the-art for many NLP tasks.

[1] Tokens are subsets of the original string, typically being individual words, subwords, characters, punctuation, etc.

[2] Parallelization enables computations to run in parallel on different cores, as opposed to computing serially on a single core. This speedup is particularly pronounced with the use of Graphical Processing Units.

[3] Bidirectional refers to the model's ability to simultaneously read previous tokens (right to left) and forward tokens (left to right).

[4] Hyperparameters dictate the network structure and training process. Unlike ordinary model parameters, they are chosen by the practitioner rather than estimated.

Although BERT and RoBERTa established a high standard in NLP, newer models have since been introduced which rival their performance on various tasks. Yang et al. (2020) at Google AI Brain Team developed XLNet which incorporates autoregressive language modeling while enabling bidirectional contexts, an extension to BERT and the Transformer model. XLNet overcomes a notable limitation of BERT, which is that its MLM method ignores dependencies between masks. XLNet highlights that the relation between masks is important, and corrupting words through masks loses those relations. The autoregressive nature of XLNet does not suffer from this limitation, and XLNet outperforms BERT on various NLP tasks.

Finally, Clark et al. (2020) introduce ELECTRA which differs from BERT primarily in its masking method. Instead of entirely masking tokens, ELECTRA instead corrupts randomized tokens with generated alternatives. In other words, the tokens are not hidden, but rather replaced. Next, ELECTRA uses a discriminative model[5] to predict whether each token is either part of the original text or generated in its pre-training. A notable advantage of ELECTRA is its computational efficiency, achieving substantially lower training times over the other methods.

To summarize, each of the four NLP text models discussed are state-of-the-art and perform exceptionally well on a variety of tasks, including sentiment analysis. The four models will be used to predict central bank speech sentiment and construct regressors that will be implemented into VIX forecasts.

[5]Discriminative models predict conditional probabilities to form a decision rule, in this case whether or not a token was generated. Conditional probabilities are often modeled using Logistic Regression, among other methods.

2 Data

This paper combines multiple data sources to perform sentiment analysis and VIX forecasting. Central bank speech data is described in subsection 2.1, data used to train the sentiment text models is covered in subsection 2.2, and the core financial and macroeconomic variables used in forecasts are highlighted in subsection 2.3. Replication instructions using these data sources and full source code are provided in section 8.

2.1 Central Bank Speech Data

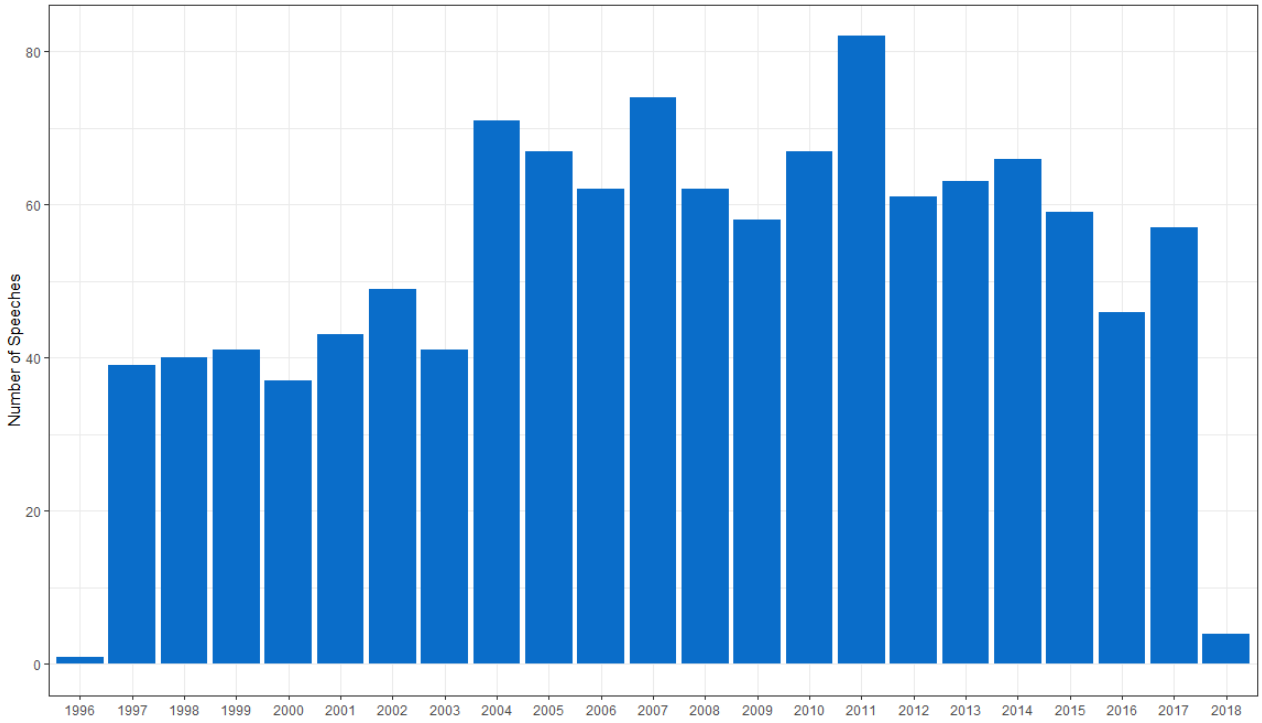
The speech text data is provided by Johnson et al. (2018) and publicly available for download at `https://osf.io/p3yr6/`. The data consists of nearly 14,000 text files, each containing the text transcript of a speech delivered by a central banker with a daily time dimension. That is, the text files do not contain any labelled sentiment predictions. Only speeches that were delivered by a central banker at the Federal Reserve are kept in order to focus solely on US monetary policy. The speeches span a total of 21 years between 19/12/1996 until 28/02/2018. The complete list of Federal Reserve speakers included in the dataset is presented in Table 1. In order to preserve the time series structure, a small number of speeches were dropped at random when there were more than 1 speech delivered on a single day, totalling 4% of the dataset being removed. The result is a dataset of 1189 speeches.

Table 1: List of Federal Reserve Speakers

Speaker	Position Title	Speech Count
Alan Greenspan	Chairman	179
Ben Bernanke	Chairman	231
Janet Yellen	Chairman	67
Jerome Powell	Chairman	43
Alice Rivlin	Vice Chairman	7
Ernest Patrikis	Vice Chairman	3
Daniel Tarullo	Board of Governors	69
Edward Gramlich	Board of Governors	13
Edward Kelley Jr	Board of Governors	13
Elizabeth Duke	Board of Governors	40
Frederic Mishkin	Board of Governors	23
Jeremy Stein	Board of Governors	11
Kevin Warsh	Board of Governors	13
Lael Brainard	Board of Governors	24
Laurence Meyer	Board of Governors	41
Mark Olson	Board of Governors	22
Randall Kroszner	Board of Governors	38
Robert Ferguson Jr	Board of Governors	74
Sarah Bloom Raskin	Board of Governors	10
Susan Schmidt Bies	Board of Governors	47
Charles Plosser	President and CEO	28
Thomas Hoenig	President and CEO	9
Timothy Geithner	President and CEO	26
William Dudley	President and CEO	91
Narayana Kocherlakota	President	14
William McDonough	President	19
Brian Sack	Executive Vice President	3
James McAndrews	Executive Vice President	7
Joseph Tracy	Executive Vice President	3
Simon Potter	Executive Vice President	21
Total		1189

The distribution of the 1189 speeches across time is displayed in Figure 2. There is a marked increase in the number of speeches starting in 2004. Mean speech count per year post-2004 was approximately 50% higher than in the pre-2004 period. In addition, the years 1996 and 2018 contain a relatively small number of speeches due to these being the start and end dates of the dataset. These attributes of the data are not expected to have important implications for their use in forecasting.

Figure 2: Number of Speeches by Year (1996-2018)



2.2 The Stanford Sentiment Treebank v2

Due to the lack of sentiment labels in the speech data, sentiment analysis must be performed to predict sentiment of every central banker's speeches. This requires the use of NLP text models which must first be trained on alternative text data that contains labelled sentiment. Developed by Socher et al. (2013), the Stanford Sentiment Treebank v2 [SST-2] is a highly influential and commonly used dataset in sentiment analysis. The core advantage of SST-2 is its large sample size ($n = 215,154$ phrases) and manually labelled sentiment, providing text models a much higher degree of predictive accuracy through better hyperparameter tuning. All NLP algorithms considered in this paper are tuned using SST-2.

2.3 Financial and Macroeconomic Data

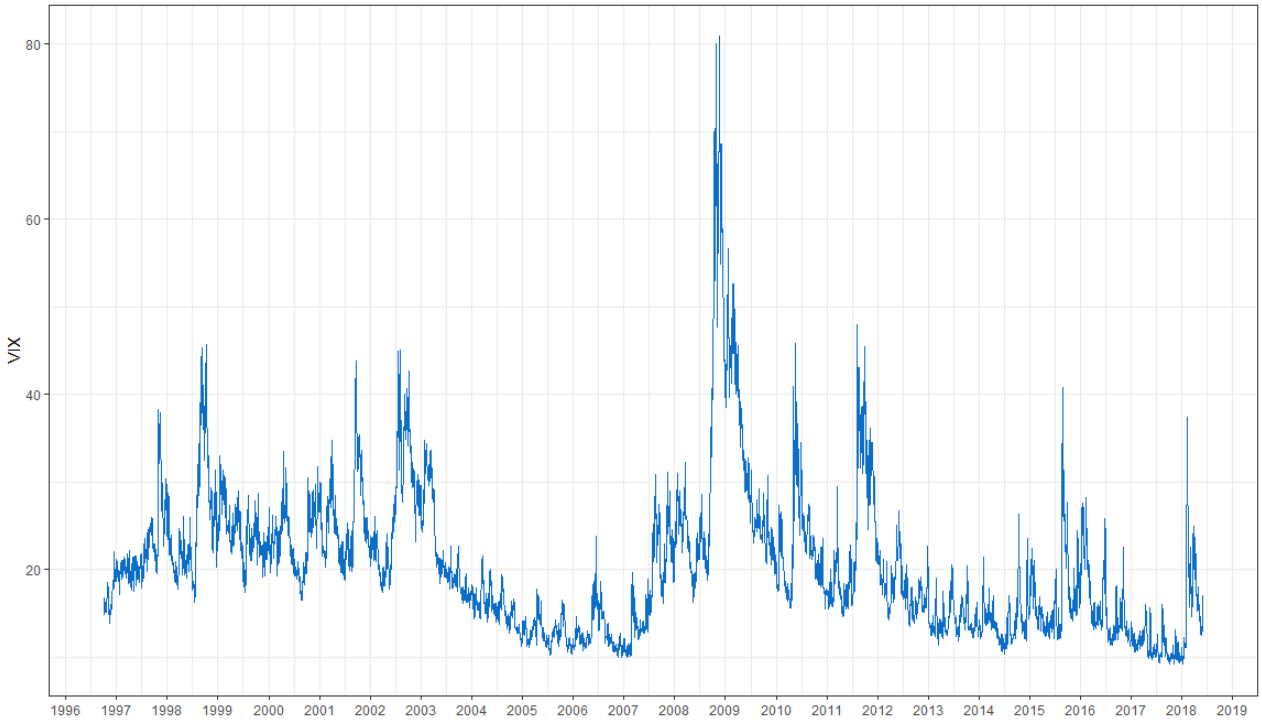
Daily financial and macroeconomic data were retrieved online. Yahoo Finance was accessed to download the VIX index [Yahoo Finance (2021a)] and SPY [Yahoo Finance (2021b)]. The SPY is a live fund tracking the S&P 500, and the retrieved data includes both the dividend-adjusted SPY level and volume. The Federal Reserve Bank of St. Louis was accessed to retried the the 10-year US treasury yield [Federal Reserve Bank of St. Louis (2021a)], the 3-month US treasury yield [Federal Reserve Bank of St. Louis (2021b)], crude oil prices [Federal Reserve Bank of St. Louis (2021c)], and Moody's 10-year Baa minus treasury credit spread [Federal Reserve Bank of St. Louis (2021d)]. Some financial and macroeconomic variables have a small number of missing observations which were imputed using the previous day's value. The summary statistics for each of these economic variables are presented in Table 2. The term spread was calculated by substracting the 3-month treasury from the 10-year treasury. Each of these variables are tested for stationarity using the Augmented Dickey-Fuller [ADF] test. The raw VIX, its log-transformation, and the volume strongly reject the null hypothesis of a unit root. On the other hand, the term spread, credit spread, and WTI oil price fail to reject the null which indicates that they are nonstationary over the sample. These variables will be transformed in subsection 4.2 before inclusion in econometric models.

Table 2: Economic Variables - Summary Statistics

Variable	Min	Max	Median	Mean	St. Dev.	Skewness	Kurtosis	ADF Test
VIX	9.14	80.86	18.98	20.37	8.37	1.93	6.61	0.01
$log(VIX)$	2.21	4.39	2.94	2.94	0.36	.47	0.16	0.01
Volume (000')	201	871026	69912	98714	101581	1.96	5.82	0.01
Term Spread (%)	-0.95	3.85	1.78	1.73	1.12	-0.17	-0.87	0.60
Credit Spread (%)	1.4	6.16	2.53	2.54	0.78	1.5	4.31	0.28
WTI Oil (USD)	10.82	145.31	50.54	55.49	29.55	0.42	-0.88	0.55

Figure 3 displays the raw (untransformed) VIX index over the sample period of 1996-2018. It has distinct behaviour around times of financial stress. For instance, it spiked during the 1997 Asian Financial Crisis, Long-Term Capital Management dissolution in 1998, the 2000 Tech Crash, 9/11, 2008 Great Financial Crisis, and 2011 Debt Ceiling Crisis. In this analysis, I will follow Lehrer et al. (2021) and Fernandes et al. (2014) and use the log transformation of the VIX, in part to contain some of its excessive skewness and kurtosis.

Figure 3: Untransformed VIX (1996-2018)



3 Sentiment Analysis Methodology

3.1 Text Pre-Processing

The text files require pre-processing before being fed into NLP algorithms. First, all text files begin with a short summary of the speaker and additional context about the speech. Since this is not part of the speaker's speech, it is systematically removed from all speeches. Additionally, a small number of tokens related to text formatting, which are not part of the speech, are embedded within the text. This requires more precision to remove; a random subset of text files were manually inspected to determine a common pattern across speeches to optimize their removal. The tokens were removed recursively to further eliminate contamination of the underlying speech. The remaining tokens are then passed to each NLP algorithm described in the next section.

3.2 NLP Text Model Sentiment Predictions

The collection of NLP text models considered in this paper is denoted by $Q = \{$BERT, RoBERTa, XLNet, ELECTRA$\}$. Raw text is first tokenized to map raw strings into distinct tokens which are identified by numerical labels. Breaking down each string increases the ability of NLP methods to capture linguistic context, particularly to capture the relation between words and punctuation. Since the NLP methods differ in their tokenization processes, every model $q \in Q$ is tokenized separately. Importantly, only the first 315[6] space-separated words of a given speech are fed to the tokenizers. All tokens following the first 315 words are unused. This approach is due to a limitation of the Transformers which are only able to take on a maximum of 512 tokens as input. A choice of 315 space-separated word count ensures each

[6] As a robustness check, other choices were also considered and do not meaningfully change the results. These include 200, 250, and 315 space-separated words at the start and end of each speech.

tokenizer will always result in no more than 512 tokens in the speech data. The limited input size is primarily due to the use of self-attention in text modeling, which requires $O(n^2)$ for both time and space complexities. Although this 315 space-separated word limitation simplifies the process of sentiment analysis, it likely reduces the precision of sentiment measures. However, the precision loss may not be large when considering that sentiment is unlikely to swing by large magnitudes throughout a speech.

Next, the tokenized text is fed as inputs into each NLP model q using the Transformers library developed by Hugging Face (2021). The library is open source and has a wide selection of pre-trained NLP models which will be superscripted with q. The text models are pre-trained on the SST-2 dataset discussed in subsection 2.2. Each NLP model q outputs a binary sentiment $s_t^q \in \{positive, negative\}$ with an associated confidence score $a_t^q \in [0, 1]$. The confidence score measures how probable the model is at predicting binary sentiment correctly. Furthermore, speeches and their associated text model output are ordered by the date that they were delivered, forming a time series. Subscript t denotes the time at which the speech was delivered. That is, the timing of a given speech and its associate computed sentiment may be used to forecast future implied volatility.

3.3 Constructing Sentiment Measurements from Text Model Output

Given the central banker speech text corpus, the goal is to get a measure of sentiment of each central banker's speech at time t to be used as a forecasting regressor in forecasting volatility at $t + h$. For modeling purposes, it is useful to define sentiment in $[0, 1]$, where 1 represents the most extreme positive sentiment and 0 the most extreme negative sentiment. The four estimated sentiment values are labeled $\tilde{x}_t^q$ for each model $q \in Q$. Although the intention is to measure sentiment $\tilde{x}_t^q$ as a real number in $[0, 1]$, $\tilde{x}_t^q$ is not the direct output of the NLP

algorithms. I map the NLP text model outputs (s_t^q, a_t^q) into $[0, 1]$ to provide a more intuitive interpretation of text sentiment. $\tilde{x}_t^q(s_t^q, a_t^q)$ is mapped as follows:

$$\tilde{x}_t^q(s_t^q, p_t^q) \equiv \begin{cases} a_t^q & \text{if } s_t^q = Positive \\ (1 - a_t^q) & \text{if } s_t^q = Negative \end{cases}$$

That is, I use the confidence scores to approximate a continuous measure of sentiment. Some caution is warranted with regards to the interpretation of $\tilde{x}_t^q(s_t^q, a_t^q)$. Although the intention is to to quantify the strength of sentiment in $[0, 1]$, the computed measure $\tilde{x}_t^q(s_t^q, a_t^q)$ differs slightly from the intended interpretation. s_t^q is a binary measure of sentiment – either positive or negative. a_t^q is a scoring measure of the binary sentiment, indicating confidence in the binary output s_t^q. My mapping is motivated by the observation that text with very strong positive or negative sentiment should be predicted more confidently than text with weaker (or ambiguous) sentiment. Therefore, this approach should be treated as an approximation to predicted sentiment rather than its definition. Sentiment is ideally defined on a continuum because a binary measure will fail to distinguish between different strengths of sentiment. Put differently, there may be information that is lost if sentiment were instead modelled as being binary. It is not feasible to quantify how well $\tilde{x}_t^q(s_t^q, a_t^q)$ measures the true underlying sentiment due to the latter being unobservable. In section 4, the predicted sentiment values $\tilde{x}_t^q(s_t^q, a_t^q)$ for each of the four models are used to construct sentiment regressors which are used in VIX forecasting models.

The summary statistics for $\tilde{x}_t^q$ are given in Table 3. Both BERT and XLNet have the largest dispersion in sentiment predictions and have relatively small absolute skewness and kurtosis relative to the other two methods. In fact, ELECTRA has notably large positive skewness

relative to the other three methods. In addition, BERT and XLNet predict less positive average sentiment values. Due to their larger variation in predicted sentiment, BERT and XLNet may better capture differences in sentiment.

Table 3: Sentiment Measurements - Summary Statistics

Variable	Min	Max	Median	Mean	St. Dev.	Skewness	Kurtosis
$\tilde{x}_t^{BERT}$	0.00	0.99	0.97	0.71	0.40	-0.93	-0.97
$\tilde{x}_t^{RoBERTa}$	0.00	0.99	0.97	0.86	0.23	-1.99	3.11
$\tilde{x}_t^{XLNet}$	0.00	0.99	0.89	0.71	0.33	-0.71	-0.92
$\tilde{x}_t^{ELECTRA}$	0.00	0.99	0.99	0.91	0.24	8.10	-3.09

Furthermore, the correlation between each sentiment measures are shown in Table 4. Each correlation is between 0.39 and 0.71, indicating a reasonably moderate-high level of consistency between NLP models.

Table 4: Measured Sentiment Correlation Matrix

	$\tilde{x}_t^{BERT}$	$\tilde{x}_t^{RoBERTa}$	$\tilde{x}_t^{XLNet}$	$\tilde{x}_t^{ELECTRA}$
$\tilde{x}_t^{BERT}$	1			
$\tilde{x}_t^{RoBERTa}$	0.63	1		
$\tilde{x}_t^{XLNet}$	0.57	0.62	1	
$\tilde{x}_t^{ELECTRA}$	0.47	0.71	0.39	1

Note: The correlation coefficient between sentiment measured by each NLP text model is estimated across the whole sample of 1189 public speeches.

4 Forecasting Framework

This section describes five forecasting models under consideration and is organized as follows. The HAR and HARX are first described in subsection 4.1 and subsection 4.2 respectively, and treated as baseline models. Next, subsection 4.3 describes a simple method to implement a single measure [SM] of averaged sentiment in a HARX context. An IV approach to dealing with measurement error is proposed in subsection 4.4 and implemented in subsection 4.5. A second

alternative measurement error strategy using factor analysis is outlined in subsection 4.6 and implemented in subsection 4.7. Finally, a test for predictive ability is described in subsection 4.8 which is used to compare empirical forecasts in section 5.

4.1 HAR

In a similar fashion to Lehrer et al. (2021) and Fernandes et al. (2014), I use the HAR model outlined by Corsi (2009) to forecast $log(VIX_t)$, denoted y_t. The HAR has become one of the more commonly used VIX forecasting models due to its simplicity and strong forecasting performance. Additionally, the HAR can easily be extended by adding covariates. For example, economic variables and sentiment measures will be implemented additively in later sections. The specification for the h-step forecasting model using the same lag index vector as Lehrer et al. (2021) of $l = (1, 5, 22)$ is given by:

$$\begin{aligned} y_t^{(k)} &= \beta_0 + \beta_1 y_t^{(1)} + \beta_5 y_t^{(5)} + \beta_{22} y_t^{(22)} + \epsilon_{t+h}^{HAR} \\ &= \beta_0 + \sum_{k \in l} \beta_k y_t^{(k)} + \epsilon_{t+h}^{HAR} \end{aligned} \tag{1}$$

where $y_t^{(k)}$ is the arthimetic mean of the logged VIX for all observations between time $t-k+1$ and t, defined as:

$$y_t^{(k)} \equiv \frac{1}{k} \sum_{j=1}^{k} y_{t-j+1} \tag{2}$$

This base HAR model will serve as the benchmark model in this paper. The additive volatility components $y_t^{(k)}$ for each $k \in l$ originate from Corsi (2009) who discusses the motivation for its structure. The HAR specification attempts to capture the behaviour of different market participants in a simple additive cascade model. Measuring mean implied volatility over 1, 5, and 22 days captures the impact of short-term, medium-term, and long-term traders

who typically rebalance their investment portfolios approximately daily, weekly, and monthly, respectively. Therefore, incorporating these three $y_t^{(k)}$ components can exploit the different trading behaviour of these three types of traders on the VIX.

4.2 HARX

The HARX, a variation of the base HAR model, is also considered. It includes the same covariates from the base HAR model with additional financial and macroeconomic variables as regressors. Inspired by Lehrer et al. (2021), Fernandes et al. (2014), and Ahoniemi (2008), I use a variety of covariates: the k-day compounded total return of the S&P 500 for each $k \in l$, the log-differenced S&P 500 volume which captures the rate of change in trading volume, the k-day compounded total return of crude oil (WTI) for each $k \in l$, the first differenced term spread between the 10-year treasury and the 3-month treasury, and the first differenced credit spread between Moody's 10-year Baa corporate and the 10-year treasury. Furthermore, Fernandes et al. (2014) raise a potentially valid concern about endogeneity, in particular between VIX and volume since both are determined by the same unobserved information set. The authors instrument their financial and macroeconomic variables (excluding return variables) with their past lags and do not find any meaningful differences in the results. Due the authors finding this issue to be fairly minor, I do not further consider potential endogeneity in these economic variables. Moreover, each transformed variable is tested for unit root with the ADF test, and all tests strongly reject the null of a unit root at the 0.001 level of signifiance.

Let z_t be the vector containing all financial and macroeconomic variables. The HARX is therefore the base HAR with z_t added as a regressor:

$$y_{t+h} = \beta_0 + \sum_{k \in l} \beta_k y_t^{(k)} + \beta_z^\top z_t + \epsilon_{t+h}^{HARX} \tag{3}$$

4.3 HARX-SM

The HARX-SM incorporates a single measure [SM] of sentiment (which inherently contains measurement error) in addition to the covariates in the HARX. In the spirit of capturing trader heterogeneity, predicted sentiment is averaged over $k \in l$ days. The motivation for averaging sentiment is similar to the motivation for the HAR itself, proposed by Corsi (2009). Short-term, medium-term, and long-term investors are trading at different time horizons. Therefore, averaging sentiment across a daily, weekly, and monthly time lengths may capture the influence of policymaker sentiment at each of these horizons. The average sentiment over the past k trading days for NLP model q will be labelled $\tilde{x}_t^{q,(k)}$ and defined as

$$\tilde{x}_t^{q,(k)} \equiv \frac{1}{k} \sum_{j=1}^{k} \tilde{x}_{t-j+1}^{q} \tag{4}$$

However, averaging values of $\tilde{x}_t^q$ poses a problem: speeches are not delivered on every day in the sample. Out of 5,449 total trading days, only 1,189 have a US central banker delivering a speech. Since $\tilde{x}_t^q$ is defined over $[0, 1]$ on days during which a speech takes place, it must also be assigned a value in $\mathbb{R}$ when a speech does not take place. For instance, a default arbitrary value of $\tilde{x}_t^q = 0$ is assigned for all t with no speeches. Although the average $\tilde{x}_t^{q,(k)}$ can now be defined, it fails to distinguish between trading days on which there is no speech and days where there is a speech whose sentiment is highly negative (i.e. $\tilde{x}_t^q \approx 0$). The ideal would be to have the sentiment regressor activate only when a speech is given, and 0 otherwise. Therefore, I propose a regressor that (1) distinguishes between positive and negative values of measured sentiment and (2) averages sentiment over the past k trading days. This regressor will be denoted as $\tilde{x}_{t,s}^{q,(k)}$ which measures model q's sentiment averaged over the past k days for sentiment s. More

precisely,

$$\tilde{x}_{t,s}^{q,(k)} \equiv \begin{cases} \frac{1}{k}\sum_{j=1}^{k} \tilde{x}_{t-j+1}^{q} & \text{if } \frac{1}{k}\sum_{j=1}^{k} \tilde{x}_{t-j+1}^{q} \geq 0.5 \text{ and } s = positive \\ (1 - \frac{1}{k}\sum_{j=1}^{k} \tilde{x}_{t-j+1}^{q}) & \text{if } \frac{1}{k}\sum_{j=1}^{k} \tilde{x}_{t-j+1}^{q} < 0.5 \text{ and } s = negative \\ 0 & \text{otherwise} \end{cases}$$

Although the sentiment regressor $\tilde{x}_{t,s}^{q,(k)}$ is defined for any $q \in Q$, I showcase the HARX-SM using sentiment derived from BERT labelled $\tilde{x}_{t,s}^{BERT,(k)}$. That is, I consider the case where q =BERT and sentiment measured by BERT is included in the model. This example is simply for the sake of clarity, three additional SM variants are performed using the same methodology with the other NLP models. Moreover, all four SM variants are tested in section 5 to ensure robustness against different sentiment measurements.

The BERT variant of the HARX-SM will be labelled the HARX-SM-BERT, which is a combination of the HARX and the mean speech sentiment $\tilde{x}_{t,s}^{BERT,(k)}$ measured by BERT, for each $s \in S$ and every $k \in l$:

$$y_{t+h} = \beta_0 + \sum_{k\in l} \beta_k y_t^{(k)} + \beta_z^{\top} z_t + \sum_{s\in S}\sum_{k\in l} \gamma_s^{BERT,(k)} \tilde{x}_{t,s}^{BERT,(k)} + \epsilon_{t+h}^{SM,BERT} \tag{5}$$

Despite large recent advances in NLP, state-of-the-art NLP models are imperfect at extracting sentiment from text due to the inherent complex structure of natural language. In other words, both the sentiment $\tilde{x}_t^q$ and its k-day average are measured with error. We should expect the HARX-SM to underperform relative to an alternative model absent of measurement error. Therefore, I propose the HARX-IV and HARX-FA, two additional models which implement different methods to deal with measurement error in sentiment.

4.4 Method 1: IV Approach to Addressing Measurement Error

The motivation behind the HARX-IV is to increase precision in the sentiment measures, leading to more accurate VIX forecasts. Before detailing the HARX-IV, I first explain a simple IV approach to dealing with measurement error in predicted sentiment.

Let $x_t \in [0, 1]$ define the true sentiment of the policymaker speech delivered at time t. The goal is to predict x_t with a consistent estimator to reduce the influence of measurement error in forecasts. The sentiment prediction $\tilde{x}_t^q$ using some NLP model $q \in Q$ at time t has measurement error u_t^q where it is assumed that $|u_t^q| > 0$, $E[u_t^q] = 0$ and $cov(u_t^q, u_t^p) = 0 \ \forall p \neq q$. This last zero covariance condition is further discussed at the end of this section due to its importance in establishing the exclusion restriction. Assuming an additive prediction error, the sentiment prediction $\tilde{x}_t^q$ for a single NLP model q is expressed as:

$$\tilde{x}_t^q = x_t + u_t^q \tag{6}$$

Because $\tilde{x}_t^q \neq x_t$, introducing $\tilde{x}_t^q$ as a regressor in a forecasting model will result in OLS being biased and inconsistent. If $cov(\tilde{x}_t^p, \tilde{x}_t^q) \neq 0$ and $cov(\tilde{x}_t^p, e_t) = 0 \ \forall p \neq q$, parameter estimation under IV will become consistent thereby improving estimation as $T \to \infty$. To illustrate, suppose a true underlying data-generating process [DGP] for y_t:

$$y_t = \beta_0 + \beta_z^\top z_t + \beta_x x_t + e_t \tag{7}$$

with some unspecified vector of non-sentiment regressors z_t. Because x_t is unobserved, it must be replaced by one of the predicted sentiments $\tilde{x}_t^q$ for some model q

$$y_t = \beta_0 + \beta_z^\top z_t + \beta_x \tilde{x}_t^q + \underbrace{(e_t - \beta_x u_t^q)}_{w_t} \tag{8}$$

Since $cov(\tilde{x}_t^q, w_t) \neq 0$, OLS yields a biased and inconsistent estimate for $\tilde{\beta}_1$ with probability limit

$$\operatorname*{plim}_{T\to\infty} \tilde{\beta}_x = \operatorname*{plim}_{T\to\infty} \frac{T^{-1} cov(y_t, \tilde{x}_t^q)}{T^{-1} var(\tilde{x}_t^q)} = \beta_x \left(\frac{var(x_t)}{var(u_t^q) + var(x_t)} \right) < \beta_x \tag{9}$$

That is, the probability limit depends on the signal-to-noise ratio. As NLP models continue to evolve and improve going forward, the noise should decline over time relative to the signal. Bias and inconsistency will remain, however, so long as sentiment is measured with non-zero noise.

For the sake of clarity, I focus the remainder of this section on instrumenting sentiment derived from BERT: $\tilde{x}_t^{BERT}$. However, any of the four NLP text model sentiments can be implemented using the same methodology. I use $\tilde{x}_t^p$ for all $p \in Q, p \neq BERT$ as instruments. That is, all NLP models other than BERT are used to instrument for sentiment predicted by BERT. Using a total of 4 NLP text models, this means 3 instruments are used in the first-stage. Moreover, the standard relevance condition and exclusion restriction must be satisfied and are discussed in the remainder of this section.

The relevance condition requires that for all $p \neq BERT$,

$$cov(\tilde{x}_t^{BERT}, \tilde{x}_t^p) \neq 0 \tag{10}$$

Table 4 previously demonstrated the correlation matrix between each of the measured sentiment values. Since every NLP model is predicting sentiment after pre-training over the same SST-2 dataset, they are all measuring the same underlying sentiment x_t engrained in the SST-2 sentiment labels. Therefore, it is expected that each measure is positively correlated with the others. Importantly, each sentiment measure has an estimated correlation with BERT somewhere between 0.47 and 0.63. This suggests that the relevance condition likely holds, although it will be further tested in first-stage results. The reduced-form for $\tilde{x}_t^{BERT}$ can additionally be established:

$$\tilde{x}_t^{BERT} = \pi_0 + \sum_{\substack{p \in Q \\ p \neq BERT}} \pi_q \tilde{x}_t^p + \pi_z^\top z_t + error \tag{11}$$

The first-stage regression output of (11) is provided in Table 5. Some moderate multicollinearity may be present due to the sentiment variables being correlated with one another, but coefficients on sentiment variables remain strongly significant regardless. In addition, the joint significance of the three sentiment instruments is strongly established using an effective F-test following Stock and Yogo (2002), with an F-stat of 328 against a threshold of 22.3. Notably, 45% of the variation in BERT-derived sentiment is captured by the first-stage[7]. The OLS fitted values from (11) will be denoted $\hat{x}_t^{BERT}$, which will be used to construct sentiment regressors in subsection 4.5 when forecasting the VIX.

[7] As a robustness check, the other non-BERT sentiment measurements are also separately treated as dependent variables, but their output is omitted from Table 5. The three other first-stage regressions yield similarly strong first-stage relevance.

Table 5: First-stage regression output of (11)

RoBERTa	0.72***
	(0.03)
XLNet	0.35***
	(0.03)
ELECTRA	0.11***
	(0.05)
All Instruments	$F = 328$
	$p = 0.00$
All Controls	$F = 1.75$
	$p = 0.11$
R^2	0.45

Note: Each first-stage regression includes all economic variables z_t as controls plus a constant. Moreover, trading days with no delivered speeches are dropped resulting in 1176 degrees of freedom. Standard errors in parentheses are computed using the Newey-West heteroskedasticity and autocorrelation consistent covariance matrix proposed by Newey and West (1987). $p < 0.01$ ***; $p < 0.05$ **; $p < 0.1$ *

Moreover, the exclusion restriction requires:

$$cov(\tilde{x}_t^p, e_t) = 0 \ \forall p \neq BERT \tag{12}$$

In other words, the predicted sentiment for each model $p \neq BERT$ must have zero covariance with the error in the DGP. This assumption is, however, difficult to test due to low power, and must instead be justified. Because the true sentiment being measured is determined by the labelled sentiment in SST-2, the deviations from the true value must be due to differences in the NLP method rather than systematic differences in true sentiment. That is, the differences observed between text model sentiment output is due to model differences, not differences in training data. Therefore, the exclusion restriction implies $cov(u_t^p, u_t^q) = 0 \ \forall p \neq q$. Although difficult to test, it is plausible that the errors are not strongly related due to the high complexity of neural networks which include millions of parameters. In addition, each NLP model's

systematic inaccuracies and limitations are different from one another, implying the resulting errors are plausibly unrelated (or only weakly related). It is still possible for a IV approach to mitigate some portion of the endogeneity associated with sentiment measurement even if there is a weak level of correlation between the errors.

4.5 HARX-IV

The HARX-IV forecasting model makes use of the first-stage results to address endogeneity in sentiment regressors. Instead of using a single measure of sentiment $\tilde{x}_t^q$ as measured by HARX-SM, HARX-IV incorporates first-stage fitted values denoted as $\hat{x}_t^q$ in subsection 4.4. Although the previous section outlined the methodology for computing first-stage fitted values for $q = \text{BERT}$, any model $q \in Q$ can be used as fitted values. If $\hat{x}_t^q$ measures true sentiment x_t more accurately than $\tilde{x}_t^q$, then HARX-IV may increase predictive ability more than HARX-SM relative to the base HAR. With the same justification proposed in the HARX-SM, the fitted values are averaged over each $k \in l$ to capture heterogeneity in traders. For some NLP model q, the k-day mean fitted sentiment is given by:

$$\hat{x}_t^{q,(k)} \equiv \frac{1}{k} \sum_{j=1}^{k} \hat{x}_{t-j+1}^q \tag{13}$$

Similar to the HARX-SM, the k-day mean fitted values are separated by positive versus negative sentiment. The motivation is identical to that outlined in subsection 4.3. The k-day mean

defined over each $s \in S$ is given by

$$\hat{x}_{t,s}^{q,(k)} \equiv \begin{cases} \frac{1}{k}\sum_{j=1}^{k} \hat{x}_{t-j+1}^{q} & \text{if } \frac{1}{k}\sum_{j=1}^{k} \hat{x}_{t-j+1}^{q} \geq 0.5 \text{ and } s = Positive \\ (1 - \frac{1}{k}\sum_{j=1}^{k} \hat{x}_{t-j+1}^{q}) & \text{if } \frac{1}{k}\sum_{j=1}^{k} \hat{x}_{t-j+1}^{q} < 0.5 \text{ and } s = Negative \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

The remainder of this section reconsiders the example of $q = \text{BERT}$ with $\hat{x}_{t,s}^{BERT,(k)}$ as sentiment regressors. That is, the HARX-IV-BERT uses BERT-derived sentiment as a dependent variable in the first-stage, and the resulting fitted values are used to compute $\hat{x}_{t,s}^{BERT,(k)}$. The HARX-IV-BERT model is therefore the HARX with added $\hat{x}_{t,s}^{BERT,(k)}$ across each $s \in S$ and every $k \in l$:

$$y_{t+h} = \beta_0 + \sum_{k\in l} \beta_k y_t^{(k)} + \beta_z^{\top} z_t + \sum_{s\in S}\sum_{k\in l} \gamma_s^{BERT,(k)} \hat{x}_{t,s}^{BERT,(k)} + \epsilon_{t+h}^{IV,BERT} \tag{15}$$

Similar to the HARX-SM, the HARX-IV has three additional variants depending on the choice of first stage method $q \in Q$. These are used as robustness checks when displaying the empirical results in section 5, and labelled HARX-IV-q when using first-stage fitted values from model q.

It should be emphasized the main difference between the HARX-SM-BERT in (5) and the HARX-IV-BERT in (15): model (5) uses $\tilde{x}_{t,s}^{BERT}$ which is the averaged predicted sentiment from BERT alone with uncorrected measurement error. This is in contrast to model (15) which uses $\hat{x}_{t,s}^{BERT}$, averaged fitted sentiment from the first stage regression using BERT as the dependent variable. Although similar in principle, the version with more accurate sentiment measurement could result in better forecast performance, assuming policymaker sentiment is

useful in VIX forecasts.

4.6 Method 2: Factor Analysis Approach to Addressing Measurement Error

A second approach to dealing with measurement error is through factor analysis. The latent sentiment can be estimated through a linear combination of measured sentiment values by dissecting variation that is common versus unique between each measure. Given sentiment data $\mathbf{X}$ with mean μ organized as:

$$\mathbf{X} = \mu + \Lambda F + error \tag{16}$$

where Λ is a matrix of factor loadings (one of which is normalized to 1) and F is a vector of latent factors. Its covariance matrix Σ can be decomposed into a linear combination between its common variation $\hat{\Lambda}\hat{\Lambda}^\top$ and remaining (unique) variation $\hat{\Psi}$ which is orthogonal to $\hat{\Lambda}\hat{\Lambda}^\top$:

$$\Sigma = \Lambda\Lambda^\top + \Psi \tag{17}$$

The unknown vector of common factors F can be estimated using both the estimated factor loading $\hat{\Lambda}$ and the estimated covariance matrix $\hat{\Sigma}$:

$$\hat{F} = \hat{\Lambda}^\top \hat{\Sigma}^{-1}\Big(\mathbf{X} - \mu\Big) \tag{18}$$

The elements of $\hat{F}$, labeled $\hat{f}_t$, are used as estimates of the factor scores f_t. Because there are four measures of sentiment and one true underlying sentiment, only one factor is estimated. The HARX-FA model is outlined outlined in subsection 4.7 where the factor scores are used as sentiment measurements using the same methodology as the other sentiment-based models.

4.7 HARX-FA

The HARX-FA integrates factor analysis scores in place of sentiment measures. Similar to the previous sentiment-based models, the HARX-FA averages these scores over $k \in l$ trading days:

$$\hat{f}_t^{(k)} \equiv \frac{1}{k} \sum_{j=1}^{k} \hat{f}_{t-j+1} \tag{19}$$

The averaged sentiment measures $\hat{f}_t^{(k)}$ for $k \in l$ are, once again, separated by sentiment $s \in S$:

$$\hat{f}_{t,s}^{(k)} \equiv \begin{cases} \frac{1}{k} \sum_{j=1}^{k} \hat{f}_{t-j+1} & \text{if } \frac{1}{k} \sum_{j=1}^{k} \hat{f}_{t-j+1} \geq 0.5 \text{ and } s = Positive \\ (1 - \frac{1}{k} \sum_{j=1}^{k} \hat{f}_{t-j+1}) & \text{if } \frac{1}{k} \sum_{j=1}^{k} \hat{f}_{t-j+1} < 0.5 \text{ and } s = Negative \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

The HARX-FA is very similar to the HARX-IV, only differing by the corrected sentiment measure. The sentiment $\hat{f}_{t,s}^{(k)}$ as predicted by factor analysis is added to the HARX, yielding the following model:

$$y_{t+h} = \beta_0 + \sum_{k \in l} \beta_k y_t^{(k)} + \beta_z^\top z_t + \sum_{s \in S} \sum_{k \in l} \gamma_s^{(k)} \hat{f}_{t,s}^{(k)} + \epsilon_{t+h}^{FA} \tag{21}$$

In essence, all three sentiment-based models are similar in concept. Although the HARX-IV and HARX-FA might be theoretical improvements over the HARX-SM, performance differences can only be settled empirically. Furthermore, the performance between the sentiment-based models should first be compared with the base HAR and HARX to test whether central bank speech sentiment adds information to VIX forecasts.

4.8 Diebold-Mariano [DM] Test for Predictive Ability

Although simple out-of-sample forecast measurements are effective in avoiding potential problems such as overfitting and data-mining results, it is useful to define a statistical test to formally compare forecasts between models. Testing forecast differences through a probabilistic lens may additionally be more effective at capturing distribution of forecast outcomes in the DGP. Diebold and Mariano (1995) outline a statistical test for testing predictive ability between two forecasts for any $h \geq 1$. The hypothesis being tested for two sets of forecast errors is:

$$H_0 : \text{Both forecasts yield equal predictive accuracy}$$

$$H_1 : \text{Not } H_0$$

In essence, H_0 is equivalent to $E[g(\epsilon_{1t}) - g(\epsilon_{2t})] = 0$ where $d = g(\epsilon_{1t}) - g(\epsilon_{2t})$ is the loss-differential. The DM test is gauging whether two sets of forecast errors, ϵ_{1t} and ϵ_{2t}, have a difference, evaluated through a function g, that is statistically significantly different from zero. Although the test is flexible towards many choices of functions g, this paper considers the absolute and squared functions, both common choices by practitioners when conducting the DM test. The mean loss differential is therefore:

$$\bar{d} = T^{-1} \sum_{t=1}^{T} [g(\epsilon_{1t}) + g(\epsilon_{2t})] \tag{22}$$

The DM test is principally an asymptotic z-test for the null that $\bar{d}$ is zero. A consistent estimate of the variance of $\bar{d}$ is given by:

$$var(\bar{d}) = \Big[\hat{\gamma}(0) + 2 \sum_{j=1}^{h-1} \hat{\gamma}(j)\Big]/T \tag{23}$$

where $\hat{\gamma}(j) = T^{-1}\sum_{t=j+1}^{T}[d_t - \bar{d}][d_{t-|j|} - \bar{d}]$ is the autocovariance function. Under H_0, the test statistic for an h-step forecast is simply $\bar{d}$ divided by the root of its variance, yielding an asymptotic standard normal distribution:

$$DM = \frac{\bar{d}}{\sqrt{\left[\hat{\gamma}(0) + 2\sum_{j=1}^{h-1}\hat{\gamma}(j)\right]/T}} \sim \mathcal{N}(0,1) \tag{24}$$

The DM test is useful because it allows for the comparison of predictive ability between different forecasts without requiring stringent (and potentially unrealistic) assumptions. Moreover, the test is robust to serial correlation, contemporaneous correlation, non-Gaussian distributions, and non-zero error means. The only required assumption is that the loss-differential is covariance stationary which will be tested before conducting inference about H_0. The test is therefore flexible in its potential applications.

5 Empirical Results

The parameters for the HAR, HARX, HARX-SM, HARX-IV and HARX-FA models are first estimated by OLS using data outlined in section 2 and associated regression output is provided in Table 6. The non-BERT variants of the SM and IV models are omitted for space considerations, but their parameter estimates are not strikingly different from BERT.

Table 6: Forecasting Model Regression Output

Model	HAR	HARX	SM-BERT	IV-BERT	FA
$Cons.$	0.0312***	0.0304***	0.0318***	0.0312***	0.0308***
	(0.0072)	(0.0071)	(0.0080)	(0.0080)	(0.0072)
$y_t^{(1)}$	0.8898***	0.9071***	0.9065***	0.9067***	0.9060***
	(0.0175)	(0.0246)	(0.0245)	(0.0245)	(0.0246)
$y_t^{(5)}$	0.0523**	0.0400	0.0411	0.0403	0.0406
	(0.0224)	(0.0281)	(0.0281)	(0.0280)	(0.0281)
$y_t^{(22)}$	0.0471***	0.0423***	0.0416***	0.0414***	0.0424***
	(0.0116)	(0.0137)	(0.0137))	(0.0136)	(0.0135)
$\tilde{x}^{BERT,(1)}_{t,positive}$			-0.0068		
			(0.0027)		
$\tilde{x}^{BERT,(5)}_{t,positive}$			0.0010		
			(0.0023)		
$\tilde{x}^{BERT,(22)}_{t,positive}$			-0.0003		
			(0.0042)		
$\tilde{x}^{BERT,(1)}_{t,negative}$			-0.0068*		
			(0.0027)		
$\tilde{x}^{BERT,(5)}_{t,negative}$			-0.0000		
			(0.0031)		
$\tilde{x}^{BERT,(22)}_{t,negative}$			0.0010		
			(0.0051)		
$\hat{x}^{BERT,(1)}_{t,positive}$				-0.0052*	
				(0.0029)	
$\hat{x}^{BERT,(5)}_{t,positive}$				-0.0005	
				(0.0025)	
$\hat{x}^{BERT,(22)}_{t,positive}$				-0.0017	
				(0.0051)	
$\hat{x}^{BERT,(1)}_{t,negative}$				-0.0048	
				(0.0065)	
$\hat{x}^{BERT,(5)}_{t,negative}$				0.0051	
				(0.0048)	
$\hat{x}^{BERT,(22)}_{t,negative}$				0.0009	
				(0.0069)	
$\hat{f}^{(1)}_{t,positive}$					-0.0122**
					0.0048
$\hat{f}^{(5)}_{t,positive}$					0.0030
					(0.0038)
$\hat{f}^{(22)}_{t,positive}$					0.0065
					(0.0054)
$\hat{f}^{(1)}_{t,negative}$					-0.0009
					(0.0016)
$\hat{f}^{(5)}_{t,negative}$					0.0013
					(0.0012)
$\hat{f}^{(22)}_{t,negative}$					0.0002
					(0.0016)
Econ Variables		$F = 0.73$ $p = 0.68$	$F = 0.77$ $p = 0.63$	$F = 0.77$ $p = 0.63$	$F = 0.73$ $p = 0.68$
Degrees of Freedom	5445	5436	5430	5430	5430
R^2	0.9670	0.9670	0.9671	0.9671	0.9671

Note: Standard errors in parentheses are computed using the Newey-West heteroskedasticity and autocorrelation consistent covariance matrix proposed by Newey and West (1987).
$p < 0.01$ ***; $p < 0.05$ **; $p < 0.1$ *

There are four variants for both the HARX-SM and HARX-IV models, separated by their different sentiment regressors derived from each NLP model $q \in Q$ and denoted by HARX-SM-q and HARX-IV-q respectively. To clarify, each variant of the HARX-SM-q uses a single measure of sentiment derived by NLP model q while the HARX-IV-q uses first-stage fitted sentiment measured by NLP model q with the remaining three sentiment measures as instruments. In addition, a fifth variant with equal-weighted [EW] sentiment measures is considered for both the SM and IV models. This fifth variant simply takes an equal weighting of each sentiment measurement across the four NLP models which are separated by each $s \in S$ and averaged over $k \in l$ days. Performance of every model is gauged through both out-of-sample forecast accuracy and through the DM test for one- (daily), five- (weekly), and twenty-two- (monthly) step forecasts. For space considerations, only the first two letters of each NLP method q are displayed. This is clarified in Table 7 where each abbreviation is shown.

Table 7: Model Abbreviations

NLP method	Abbreviation
BERT	BE
RoBERTa	RO
XLNet	XL
ELECTRA	EL
Equal-Weight	EW

The out-of-sample forecast accuracy based on four simple diagnostics. First, the Mean Forecast Error [MFE] measures average differences between actual values (y_t) and forecasts($\hat{y}_t$):

$$MFE = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \hat{y}_t\right) \tag{25}$$

Since positive and negative errors tend to offset each other, the MFE is not intended to gauge forecast accuracy. Instead, it measures bias in out-of-sample forecasts – either systematically

above or below the true values. Forecast accuracy is instead computed through the remaining three diagnostics. The Mean Square Error [MSE] and Mean Absolute Error [MAE] measure average square and absolute differences in errors, defined as

$$MSE = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \hat{y}_t\right)^2 \tag{26}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|y_t - \hat{y}_t\right| \tag{27}$$

Finally, the Mean Absolute Percent Error [MAPE] is the average absolute deviation between forecasts and actual values, measured in percent units:

$$MAPE = \frac{100}{T}\sum_{t=1}^{T}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \tag{28}$$

Table 8 displays the four out-of-sample diagnostics for every model over each $h = 1, 5, 22$ forecast horizons. The MFE is virtually zero for all models which implies little forecast bias. For all forecast horizons, MFE falls in absolute terms under the sentiment-based models relative to both the HAR and HARX. The HAR minimzes the MSE across all horizons, including against the non-sentiment HARX model. The MAE and MAPE tell a slightly different story. A number of sentiment-based models outperform based on MAE and MAPE, but not all. For instance, the SM-BE outperforms the HAR and HARX across all horizons based on the MAPE. Some additional variants increase out-of-sample accuracy under both the MAE and MAPE depending on the forecast horizon. Moreover, there is no noticeable difference or pattern between the SM and IV models. The FA method achieves marginal outperformance over the HAR and HARX over medium- and long-term horizons under the MAPE.

However, it should be highlighted that the performance differences in out-of-sample accuracy

is fairly small in magnitude across all models. Overall, these results are not surprising and appear to be consistent with Fernandes et al. (2014) who finds similar out-of-sample performance measured across a different set of HAR-based models. The researchers additionally mention that the VIX's highly persistent nature implies that the base HAR is particularly well-suited in forecasting implied volatility, suggesting outperformance to be a challenge.

Table 8: Out of sample VIX Forecasts

	MFE	MSE	MAE	MAPE
One-Step Ahead				
HAR	0.002 68	0.00488	0.049 35	1.719 27
HARX	0.002 66	0.004 90	0.049 37	1.719 93
SM-BE	0.002 57	0.004 90	0.049 33	1.718 28
SM-RO	0.002 62	0.004 89	0.049 36	1.719 67
SM-XL	0.002 62	0.004 90	0.049 39	1.720 62
SM-EL	0.002 66	0.004 90	0.049 35	1.719 19
SM-EW	0.002 64	0.004 90	0.049 35	1.719 22
IV-BE	0.002 64	0.004 89	0.049 35	1.719 17
IV-RO	0.002 57	0.004 90	0.049 37	1.719 95
IV-XL	0.00255	0.004 89	0.04932	1.71831
IV-EL	0.00255	0.004 90	0.049 37	1.720 01
IV-EW	0.002 60	0.004 89	0.049 34	1.718 97
FA	0.002 57	0.004 90	0.049 36	1.719 63
Five-Steps Ahead				
HAR	0.011 80	0.01771	0.097 64	3.413 86
HARX	0.011 17	0.017 75	0.097 49	3.406 80
SM-BE	0.010 26	0.017 76	0.097 40	3.403 35
SM-RO	0.010 58	0.017 77	0.097 51	3.407 66
SM-XL	0.010 81	0.017 78	0.097 56	3.409 48
SM-EL	0.010 38	0.017 75	0.09743	3.405 27
SM-EW	0.010 60	0.017 77	0.09743	3.405 32
IV-BE	0.010 80	0.017 76	0.097 48	3.407 49
IV-RO	0.010 29	0.017 77	0.097 49	3.407 11
IV-XL	0.010 34	0.017 76	0.09743	3.40483
IV-EL	0.01018	0.017 76	0.097 44	3.405 00
IV-EW	0.010 54	0.017 77	0.097 46	3.406 28
FA	0.010 66	0.017 76	0.097 60	3.412 01
Twenty-Two-Steps Ahead				
HAR	0.039 59	0.04778	0.16630	5.830 24
HARX	0.037 35	0.048 60	0.167 07	5.835 32
SM-BE	0.035 21	0.048 72	0.166 80	5.822 82
SM-RO	0.035 97	0.048 89	0.167 38	5.843 16
SM-XL	0.036 19	0.049 49	0.168 01	5.868 32
SM-EL	0.035 62	0.048 60	0.166 54	5.81730
SM-EW	0.036 02	0.048 98	0.167 40	5.844 99
IV-BE	0.035 74	0.049 28	0.167 74	5.854 87
IV-RO	0.035 51	0.048 79	0.167 01	5.831 05
IV-XL	0.035 60	0.049 08	0.167 32	5.840 19
IV-EL	0.035 19	0.048 78	0.166 86	5.825 38
IV-EW	0.035 61	0.049 08	0.167 45	5.844 80
FA	0.03488	0.049 00	0.166 91	5.826 41

Note: The table displays out-of-sample forecast diagnostics assessed through cross-validation. A rolling window starting with the first 30% of the sample is used to estimate the parameters in each model and predict y_{t+h}. The MFE, MSE, MAE, and MAPE are computed using the h-step out-of-sample forecasts. The best performing model for every diagnostic is underlined in each panel.

The DM test is considered to be the primary formal test in this paper to gauge predictive ability between forecasts. The test, as outlined in subsection 4.8, crucially depends on stationarity

of every loss-differential being tested. The loss-differential for each pair of forecast errors are tested with the ADF test and the unit root null hypothesis is strongly rejected for all pairs at the 1% level of significance.

Table 9 highlights the p-values in the DM test using the MSE as a loss function. First, I cannot reject the null that the HARX and HAR achieve the same predictive ability for short- and long-term forecasts, although the HARX does marginally reject the DM null hypothesis over a medium-term horizons at the 10% level of significance. Each of the three sentiment-based models (SM, IV, FA) either strongly or marginally reject the null against the HAR for one day forecasts. Over one- and twenty-two day horizons, all of the sentiment-based models achieve marginally significant outperformance against the HAR with the exception of the SM-XL and SM-BE which are wholly insignificant. None of the sentiment-based models are found to outperform the HARX regardless of the forecast horizon.

Although mostly consitent with the MSE table, the DM test using the MAE does find a few different results. Table 10 displays the p-values for the MAE version of the DM test. The HARX and every sentiment-based model outperforms the HAR with p-values very close to zero over one- and five-step forecasts. All models are much less conclusive over twenty-two day forecasts, however. Sentiment-based models additionally outperform the HARX with fairly high confidence over one-step forecasts. This is not true, however, over five- and twenty-two-step horizons which are entirely insignificant.

Although the results of the DM test do vary between the MAE and MSE, a few conclusions are consistent between both tables. First, sentiment-based models significantly outperform the HAR over short- and medium-term forecasts. Long-term forecasts are less clear, with the MSE variant of the DM test finding marginal significance across a few models, while the MAE

variant finds only the IV-RO forecasts marginally significant against the HAR. Furthermore, all sentiment-based models significantly outperform the HARX over one-step horizons under the MAE except for the SM-RO. This result does not hold under the MSE where no model significantly outperforms against the HARX. Another important result is that the FA model yields largely similar results compared to the SM and IV. That is, its significance largely follows that of the SM and IV variants.

Table 9: Diebold-Mariano Test for MSE

	HAR	HARX	SM-BE	SM-RO	SM-XL	SM-EL	SM-EW	IV-BE	IV-RO	IV-XL	IV-EL	IV-EW
One Step Ahead												
HARX	0.207											
SM-BE	0.071*	0.178										
SM-RO	0.048**	0.142	0.592									
SM-XL	0.108	0.323	0.403	0.254								
SM-EL	0.078*	0.246	0.928	0.400	0.534							
SM-EW	0.083*	0.246	0.674	0.304	0.472	0.782						
IV-BE	0.084*	0.249	0.807	0.248	0.509	0.864	0.867					
IV-RO	0.076*	0.238	0.971	0.423	0.461	0.847	0.700	0.787				
IV-XL	0.068*	0.204	0.987	0.366	0.388	0.877	0.552	0.524	0.940			
IV-EL	0.045**	0.159	0.691	0.748	0.295	0.491	0.382	0.392	0.532	0.548		
IV-EW	0.068*	0.211	0.953	0.306	0.394	0.786	0.515	0.403	0.859	0.889	0.501	
FA	0.065*	0.167	0.839	0.662	0.407	0.801	0.613	0.651	0.842	0.853	0.795	0.900
Five Steps Ahead												
HARX	0.051											
SM-BE	0.022**	0.220										
SM-RO	0.030**	0.223	0.568									
SM-XL	0.030**	0.243	0.955	0.630								
SM-EL	0.035**	0.212	0.441	0.586	0.494							
SM-EW	0.034**	0.248	0.676	0.703	0.732	0.467						
IV-BE	0.031**	0.209	0.467	0.728	0.496	0.878	0.454					
IV-RO	0.037**	0.227	0.470	0.590	0.491	0.952	0.488	0.912				
IV-XL	0.021**	0.187	0.385	0.679	0.506	0.958	0.481	0.879	0.986			
IV-EL	0.022**	0.140	0.310	0.302	0.327	0.616	0.305	0.615	0.571	0.701		
IV-EW	0.028**	0.190	0.356	0.404	0.404	0.827	0.294	0.420	0.782	0.631	0.825	
FA	0.044**	0.241	0.604	0.886	0.624	0.894	0.742	0.971	0.921	0.925	0.698	0.800
Twenty-Two Steps Ahead												
HARX	0.192											
SM-BE	0.107	0.278										
SM-RO	0.095*	0.247	0.975									
SM-XL	0.078*	0.311	0.570	0.656								
SM-EL	0.047**	0.120	0.388	0.083	0.253							
SM-EW	0.091*	0.283	0.983	0.982	0.524	0.244						
IV-BE	0.079*	0.324	0.519	0.599	0.845	0.226	0.494					
IV-RO	0.061*	0.137	0.301	0.104	0.260	0.561	0.212	0.245				
IV-XL	0.078*	0.270	0.659	0.772	0.625	0.257	0.654	0.446	0.255			
IV-EL	0.082*	0.227	0.684	0.601	0.487	0.483	0.664	0.429	0.358	0.516		
IV-EW	0.081*	0.267	0.842	0.863	0.576	0.209	0.753	0.481	0.209	0.688	0.532	
FA	0.064*	0.297	0.543	0.536	0.779	0.202	0.516	0.815	0.235	0.560	0.408	0.521

Note: Table displays the DM test p-values for the null hypothesis that the row and column forecasts yield identical predictive ability using the MSE loss function.
$p < 0.01$ ***; $p < 0.05$ **; $p < 0.1$ *

Table 10: Diebold-Mariano Test for MAE

	HAR	HARX	SM-BE	SM-RO	SM-XL	SM-EL	SM-EW	IV-BE	IV-RO	IV-XL	IV-EL	IV-EW
One Step Ahead												
SM-BE	0.000***	0.007***										
SM-RO	0.002***	0.138	0.473									
SM-XL	0.000***	0.037**	0.083*	0.777								
SM-EL	0.000***	0.049**	0.390	0.947	0.614							
SM-EW	0.000***	0.020**	0.291	0.795	0.207	0.786						
IV-BE	0.000***	0.033**	0.318	0.881	0.445	0.933	0.768					
IV-RO	0.000***	0.043**	0.425	0.858	0.463	0.762	0.907	0.941				
IV-XL	0.000***	0.040**	0.400	0.776	0.432	0.818	0.986	0.765	0.927			
IV-EL	0.001***	0.091*	0.507	0.834	0.648	0.925	0.904	0.981	0.972	0.912		
IV-EW	0.000***	0.042**	0.474	0.708	0.432	0.754	0.983	0.685	0.885	0.954	0.864	
FA	0.000***	0.033**	0.615	0.713	0.418	0.729	0.816	0.704	0.797	0.812	0.797	0.839
Five Steps Ahead												
HARX	0.008											
SM-BE	0.007***	0.447										
SM-RO	0.009***	0.445	0.788									
SM-XL	0.006***	0.345	0.808	0.958								
SM-EL	0.008***	0.293	0.469	0.438	0.600							
SM-EW	0.004***	0.244	0.343	0.588	0.556	0.788						
IV-BE	0.016**	0.548	0.874	0.919	0.982	0.545	0.556					
IV-RO	0.007***	0.284	0.462	0.473	0.578	0.936	0.812	0.580				
IV-XL	0.020**	0.628	0.974	0.830	0.905	0.550	0.553	0.851	0.556			
IV-EL	0.016**	0.430	0.653	0.678	0.753	0.867	0.956	0.712	0.904	0.667		
IV-EW	0.016**	0.469	0.694	0.778	0.833	0.739	0.885	0.506	0.763	0.457	0.871	
FA	0.033**	0.819	0.804	0.673	0.691	0.457	0.496	0.721	0.486	0.813	0.573	0.619
Twenty-Two Steps Ahead												
HARX	0.217											
SM-BE	0.224	0.477										
SM-RO	0.237	0.517	0.875									
SM-XL	0.165	0.419	0.854	0.988								
SM-EL	0.120	0.268	0.482	0.132	0.410							
SM-EW	0.199	0.443	0.862	0.643	0.739	0.362						
IV-BE	0.154	0.378	0.913	0.933	0.838	0.419	0.795					
IV-RO	0.080	0.190	0.258	0.032	0.260	0.404	0.138	0.263				
IV-XL	0.159	0.374	0.878	0.736	0.639	0.451	0.952	0.680	0.246			
IV-EL	0.161	0.366	0.745	0.578	0.655	0.547	0.831	0.673	0.286	0.792		
IV-EW	0.164	0.382	0.779	0.500	0.623	0.415	0.824	0.656	0.181	0.759	0.883	
FA	0.136	0.331	0.846	0.955	0.922	0.378	0.744	0.848	0.260	0.702	0.662	0.660

Note: Table displays the DM test p-values for the null hypothesis that the row and column forecasts yield identical predictive ability using the MAE loss function.
$p < 0.01$ ***; $p < 0.05$ **; $p < 0.1$ *

6 Discussion

In short, the evidence presented in this paper suggests that sentiment measured from Federal Reserve speeches can be useful in implied volatility forecasts, although the magnitude of performance differences in how researchers incorporate sentiment is not large. All sentiment-based models find some success in outperforming the HAR over shorter forecast horizons under both the DM test and out-of-sample accuracy. In addition, policymaker sentiment is more effective

over shorter forecast horizons and becomes less effective as the horizons expands. Sentiment-based models also find some limited outperformance against the HARX, although this result does depend on the choice of test diagnostic. Overall, sentiment-based models appear somewhat more effective when using MAE for both out-of-sample accuracy and under the DM test. Squaring the errors does erode a portion of the outperformance, however. Moreover, the FA and IV methods to address measurement error yield similar results to the single sentiment measures. The equal-weight variants of the HARX-SM and HARX-IV do not appear to be meaningful improvements across any test or diagnostic.

Practitioners may find the results presented in this apper useful for a few reasons. The use of policymaker sentiment can reduce VIX forecast error, particularly when performing nowcasts and short-term forecasts. For instance, institutions sensitive to market risks and volatility may benefit from more precise short-term VIX modeling. In addition, the two measurement error approaches may be useful as data becomes more widely available, especially due to the reliance on asymptotic theory in the IV and FA approaches.

The methodology used in this paper does face a variety of limitations, however. Avenues for future research are proposed for each limitation. First, the definition of measured sentiment is based on confidence scores rather than binary positive/negative algorithm output. Since this is a mere approximation to sentiment, it would be beneficial if future NLP methods could more effectively predict sentiment on a continuum rather than being binary. Additionally, restricting sentiment to one dimension – positive versus negative – may be inferior to a more varied set of possible values. Some NLP methods do exist for multidimensional sentiment, but they typically involve smaller training sample sizes and have less model development. Sentiment itself, whether measured in one or multiple dimensions, may also have different impacts on

financial markets depending on the type of policymaker communication under consideration. For example, a speech discussing general macroeconomic research may have different impacts on financial markets compared to an alternative speech announcing new policy changes. Conditioning on the type of speech is likely important, and implementing a topic model to address this issue may be an effective approach.

Moreover, only the first 315 space-separated words of a given speech are used to generalize the sentiment of the entire speech. This is due to the self-attention mechanism of Transformers-based models requiring $O(n^2)$ time and memory complexities. Beltagy et al. (2020) propose the longformer, a new self-attention mechanism with time and memory complexities of $O(n)$. The linear scaling allows it to take a much larger number of tokens with minimal computational power, even for files with thousands of tokens. Implementing the longformer could therefore allow for a large token input which would effectively increase precision of sentiment measurement.

Finally, more research is needed in a sentiment analysis context to further test the IV and FA approaches for reducing measurement error, both against each other and against models using single measures of sentiment. Both approaches may additionally require a larger sample size to detect small forecast improvements, largely because sentiment variables themselves are usually not the most critical variables in econometric models. The type of sentiment may be relevant as well, and performing additional tests on consumer and investor sentiment might yield more pronounced results.

7 Conclusion

This paper uses four NLP text models to predict sentiment of 1189 central bank policymaker speeches. These are used to construct sentiment regressors implemented in VIX forecasting models with a sample of 5449 trading days. A total of five predictive models were built, all based on the HAR proposed by Corsi (2009), three of which are sentiment-based. The first sentiment implementation method is to include a single measure of sentiment, while the other two attempt to correct measurement error through instrumental variable and factor analysis approaches. All sentiment-based models average sentiment over different time lengths to capture trader heterogeneity.

I establish a few notable results based on both out-of-sample forecast accuracy and the DM test for predictive ability which can be summarized as follows. The joint inclusion of policymaker sentiment with financial and macroeconomic variables in the HAR model significantly improves predictive ability across short- and medium-term forecast horizons through the DM test. Some variants of the sentiment-based models additionally find outperformance in out-of-sample tests, although the magnitude of the differences are mostly small. Depending on the chosen test diagnostic, sentiment-based models can sometimes outperform the HARX under the MAE. Moreover, sentiment is most effective over the shortest time horizon and becomes less effective as the horizon expands. In short, these findings suggest that policymaker sentiment may be useful for nowcasts and short-term forecasts, although the potential performance gains are not large.

In addition, the IV approach to reduce measurement error in predicted sentiment achieves a strong first-stage ($F = 328$ for instrument joint significance test and $R^2 = 0.45$) – a result which

may be useful for future sentiment measurement error research. With this application, there is no optimal strategy to address measurement error in sentiment variables and neither the IV nor the FA methods achieve significant outperformance against the single sentiment measure models. Future research should extend this investigation to other economic applications where sentiment measured from text data is utilized as an explanatory variable.

8 Replication

The full Python and R source code are on my GitHub: `https://github.com/nvanhell/MA-Research-Project`. A base directory should be defined where both the Python script and R code are to be placed. All folder paths described in this section are relative to the base directory. Details to fully replicate this project are given in two parts, one for the NLP text modeling and another for the forecasting.

First, NLP text models from Hugging Face (2021) were used to construct sentiment measurements in Python. The text data was downloaded from Johnson et al. (2018) at `https://osf.io/p3yr6/`. The raw txt files should be placed in the following relative path: /source/txt/. The text manual that is included in the download should be placed in /source/ – it contains important metadata about each speech. After executing the Python script, a csv file will be generated in /data_files_generated/generated_speech_sentiment.csv which contain predicted sentiment for each speech and for every NLP text model. This file will form the basis of the forecasting section. The Python file will take about half a day to run (depending on CPU) without parallelization due to the high computational cost of the NLP algorithms and text pre-processing.

Second, the forecasting section requires additional datasources from the Federal Reserve Bank of St. Louis and Yahoo Finance. These are described in subsection 2.3 and which includes sources for each data file. Sources should be individually downloaded and placed in a /source/volatility folder relative to the base directory. The file names should be edited to match those in the R file. All output used in this paper will be generated after executing the R file.

References

Ahoniemi, K. (2008). Modeling and Forecasting Implied Volatility – an Econometric Analysis of the VIX Index. *Helsinki Center of Economic Research* 129. DOI: https://dx.doi.org/10.2139/ssrn.1033812.

Beltagy, I., M. Peters, and A. Cohan (2020). Longformer: The Long-Document Transformer. *ArXiv*. URL: https://arxiv.org/abs/2004.05150.

Bernanke, B. and K. Kuttner (2005). What Explains the Stock Market's Reaction to Federal Reserve Policy? *The Journal of Finance* 60.3, pp. 1221–1257. DOI: https://doi.org/10.1111/j.1540-6261.2005.00760.x.

Chicago Board Options Exchange (2021). *Cboe Volatility Index*. URL: https://cdn.cboe.com/resources/vix/vixwhite.pdf (visited on 07/03/2021).

Clark, K. et al. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *ArXiv* abs/2003.10555. URL: https://arxiv.org/abs/2003.10555.

Clements, A. and J. Fuller (2012). Forecasting increases in the VIX: A time-varying long volatility hedge for equities. *NCER Working Paper Series* 88.

Corsi, F. (2009). A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Fianncial Econometrics* 7.2, pp. 174–196. DOI: https://doi.org/10.1093/jjfinec/nbp001.

Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*. URL: https://arxiv.org/abs/1810.04805v2.

Diebold, F. and R. Mariano (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics* 13, pp. 253–265.

Federal Reserve Bank of St. Louis (2021a). *10-Year Treasury Constant Maturity Rate*. URL: https://fred.stlouisfed.org/series/DGS10 (visited on 07/09/2021).

— (2021b). *3-Month Treasury Constant Maturity Rate*. URL: https://fred.stlouisfed.org/series/DGS3MO (visited on 07/09/2021).

— (2021c). *Crude Oil Prices: West Texas Intermediate (WTI) - Cushing, Oklahoma*. URL: https://fred.stlouisfed.org/series/DCOILWTICO (visited on 07/09/2021).

— (2021d). *Moody's Seasoned Baa Corporate Bond Yield Relative to Yield on 10-Year Treasury Constant Maturity*. URL: https://fred.stlouisfed.org/series/BAA10Y (visited on 07/09/2021).

Fernandes, M., M. Medeiros, and M. Scharth (2014). Modeling and predicting the CBOE market volatility index. *Journal of Banking & Finance* 40, pp. 1–10. DOI: https://doi.org/10.1016/j.jbankfin.2013.11.004.

Gentzkow, M., B. Kelly, and M. Taddy (2017). Text as Data. *National Bureau of Economic Research* 23276. DOI: 10.3386/w23276.

Hugging Face (2021). URL: https://huggingface.co/ (visited on 07/08/2021).

Jiang, Y. and E. Lazar (2020). Forecasting VIX Using Filtered Historical Simulation. *Journal of Financial Econometrics*. DOI: https://doi.org/10.1093/jjfinec/nbaa041.

Johnson, J., V. Arel-Bundock, and V. Portniaguine (2018). Adding rooms onto a house we love: Central banking after the Global Financial Crisis. *Public Administration* 97.3, pp. 546–560. DOI: https://doi.org/10.1111/padm.12567.

Keynes, J. M. (1936). The General Theory of Employment, Interest and Money. *Palgrave Macmillan*.

Lehrer, S., T. Xie, and X. Zhang (2021). Social Media Sentiment, Model Uncertainty, and Volatility Forecasting. *Economic Modelling* 102, p. 105556. ISSN: 0264-9993. DOI: https://doi.org/10.1016/j.econmod.2021.105556.

Liu, Y. et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*. URL: https://arxiv.org/abs/1907.11692v1.

Newey, W. and K. West (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelationconsistent Covariance Matrix. *Econometrica* 55.3, pp. 703–708. DOI: https://doi.org/10.2307/1913610.

Shvimer, Y., V. Murinde, and A. Herbon (2021). Forecasting the CBOE VIX with a hybrid LSTM-ARIMA model and sentiment analysis. *Centre for Global Finance Working Paper Series* 1.

Socher, R. et al. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *EMNLP* 1631, pp. 1631–1642.

Stock, J. and M. Yogo (2002). Testing for Weak Instruments in Linear IV Regression. *National Bureau of Economic Research* 284. DOI: 10.3386/t0284.

Vähämaa, S. and J. Äijö (2010). The Fed's policy decisions and implied volatility. *The Journal of Futures Markets* 31.10, pp. 995–1010. DOI: https://doi.org/10.1002/fut.20503.

Vaswani, A. et al. (2017). Attention Is All You Need. *ArXiv*. URL: https://arxiv.org/abs/1706.03762.

Wang, Y-H., A. Keswani, and S. Taylor (2005). The Relationships between Sentiment, Returns and Volatility. *International Journal of Forecasting* 21.1. DOI: https://doi.org/10.1016/j.ijforecast.2005.04.019.

Yahoo Finance (2021a). *CBOE Volatility Index (VIX)*. URL: https://finance.yahoo.com/quote/%5C%5EVIX?p=%5C%5EVIX (visited on 07/09/2021).

— (2021b). *SPDR S&P 500 ETF Trust (SPY)*. URL: https://finance.yahoo.com/quote/SPY?p=SPY (visited on 07/09/2021).

Yang, Z. et al. (2020). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *ArXiv*. URL: https://arxiv.org/abs/1906.08237v2.

Young Chang, B. and B. Feunou (2014). Measuring Uncertainty in Monetary Policy Using Realized and Implied Volatility. *Bank of Canada Working Paper*.

Zhu, Y. et al. (2015). Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. *ArXiv*. URL: https://arxiv.org/abs/1506.06724.