

Textual Regression for Realized Volatility: A Model for Long-Term Forecasting

Navid Parvini*, Hirbod Assa[†]

*Kent Business School, Canterbury, UK. N.Parvini@kent.ac.uk

[†]Modellibrary, UK. Assa.Hirbod@gmail.com

Corresponding author: Navid Parvini (N.Parvini@kent.ac.uk)

Textual Regression for Realized Volatility: A Model for Long-Term Forecasting

Abstract

This study investigates the role of textual information in forecasting realized volatility in financial markets. Using state-of-the-art large language models (LLMs) such as LLaMA, OPT, BERT, and RoBERTa, we develop novel textual regression models to predict one-day-, one-week-, two-week-, and one-month-ahead volatility for corn, soybean, and wheat markets based on news articles published by media agencies. Our results show that these models significantly outperform autoregressive models like HAR, as well as sentiment-based approaches, particularly in long-term forecasting. The findings also reveal relationships between forecasting accuracy and both the size and type of the language model. Shapley value analysis demonstrates that market-related terms enhance short-term forecasts, while production- and weather-related terms drive long-term predictions. Our findings underscore the utility of textual data for forecasting volatility and provide a foundation for further applications in financial markets and risk management.

Keywords: Volatility modelling, Text regression, Natural language processing (NLP), Large language models (LLM), Agriculture

Word count: 6991

1. Introduction

Effective risk management hinges on the ability to measure and forecast risk accurately. As financial markets are inherently uncertain, the capacity to anticipate shifts in risk conditions is critical for investors seeking to optimize returns, traders operating under defined constraints, and portfolio managers aiming to achieve balanced exposure. Consequently, financial market players actively track shifts in risk conditions, dynamically adjusting their portfolios—scaling back exposures during periods of heightened risk and expanding them when risk subsides. Among the various dimensions of risk, volatility plays a central role, serving as a proxy for uncertainty and a key determinant of asset allocation, pricing, and risk-adjusted returns.

In recent decades, advancements in volatility forecasting have been shaped by the evolving availability of financial data and methodological innovation. While early studies leveraged parametric models like GARCH (Engle, 1982; Bollerslev, 1986) and stochastic volatility frameworks (Taylor, 1982), the advent of high-frequency intraday data has shifted attention to realized volatility measures, which capture price variability with greater precision (Andersen and Bollerslev, 1998; Andersen et al., 2001; Barndorff-Nielsen and Shephard, 2002). Building on this, models like the Heterogeneous Autoregressive (HAR) model of Corsi (2008) have become prominent for their simplicity and empirical success, outperforming traditional approaches. These developments underscore the growing importance of accurate volatility forecasting in enhancing risk management and investment decision-making strategies. While these models effectively capture volatility clustering, they often overlook the influence of exogenous factors. Despite introducing models like HAR-X and GARCH-X that incorporate exogenous factors, integrating such exogenous information into forecasting frameworks remains a significant methodological hurdle, demanding innovative approaches to disentangle and leverage these complex drivers.

In recent years, financial research has increasingly turned to textual data from news outlets, regulatory filings, and social media. However, the integration of text mining into economic modelling remains in its early stages. Many existing studies rely on simplistic text representations, such as dictionary-based sentiment measures or bag-of-words techniques, which often fail to capture the nuanced nature of textual information.

The underutilization of textual data in economic research stems from significant challenges. Unlike structured numerical data, text is inherently complex and nuanced, requiring sophisticated analytical models to extract meaningful insights. Unfortunately, these models are inaccessible to many researchers due to their steep technical demands and substantial computational costs. As a result, much of the potential of text analysis in finance remains untapped. Despite these barriers, the field holds immense promise. The current limitations suggest that text analysis in finance is only scratching the surface, with future research poised to explore broader text sources and develop more advanced techniques for extracting actionable information. Motivated by this, we seek to contribute to this emerging field by employing advanced text representations built with large language models (LLMs) to enhance predictions of volatility.

The present study leverages this information-rich resource to explore the relationship between media news text and the future state of agricultural commodities’ volatilities. Media outlets play a critical role in disseminating market-relevant information to stakeholders.

Their goal is to present this information in a clear and accessible manner, making it a valuable source for understanding market conditions. Unlike individual stocks, which can be influenced by individual opinions published on social media—making it challenging to identify effective information sources—the commodities market, particularly agricultural commodities, tends to follow macro-level information related to production, trade, policy, economic news, weather, and day-to-day market movements. By utilizing media news, we aim to capture the primary information sources that potentially drive market fluctuations.

In the context of agricultural commodities, several studies have explored the application of autoregressive models like HAR and GARCH to forecast volatility (Degiannakis et al., 2022; Luo et al., 2022; Sanjuán-López and Dawson, 2017; Dawson, 2014; Karali and Power, 2013). While these models effectively capture volatility clustering, they often overlook the influence of exogenous factors such as policy announcements, weather patterns, and macroeconomic uncertainties. Recognizing these limitations, recent research has sought to incorporate external variables, such as USDA reports (Bian et al., 2022a; Indriawan et al., 2021; Cao and Robe, 2021), economic policy uncertainty (Xiang and Borjigin, 2024; Dutta et al., 2024), trade tensions (Berger et al., 2021), and weather shocks (Gallic and Vermandel, 2020), among others. However, these efforts have faced challenges due to the sheer abundance and complexity of factors influencing agricultural price volatility. Mustafa et al. (2023) provides a systematic review of the volatility modelling in agricultural market, the datasets used, and the outcomes.

This paper aims to fill this gap by proposing a textual regression framework that directly uses textual information to forecast volatility. The central premise is that autoregressive models, while useful, may fall short in capturing the full range of market-relevant information—particularly qualitative signals embedded in text. Despite the richness of textual data sources such as news, their application in volatility modeling remains relatively untapped. Our paper is the first to use LLMs in volatility regression. This study is close to Chen et al. (2022) in that we use sophisticated LLM methods to quantify the information in text. However, despite Chen et al. (2022) where they classify the return direction, we forecast volatility using a regression approach. To contextualize our contribution, we first examine the prevailing methodologies in financial text mining, which predominantly use supervised machine learning tailored to specific tasks. Two general approaches dominate this field. The first involves extracting a one-dimensional index from textual data or reducing high-dimensional text representations into a single dimension. For instance, Engle et al. (2020) propose a text-based methodology for hedging climate risk using information from media news. In their study, they develop a news-based climate index by counting climate-related words in media reports. Similarly, Baker et al. (2016) construct an economic policy uncertainty index by quantifying the occurrence of terms in news articles that suggest economic uncertainty.

Regression models using sentiment indices also fall into this category, where high-dimensional textual vectors are condensed into one-dimensional sentiment measures. Traditional approaches typically rely on dictionary-based methods, which offer the advantage of interpretability. Notable example include the study by Shapiro et al. (2022) which adopt this framework. More recent research leverages large language models (LLMs), such as Huang et al. (2023), where BERT model is fine-tuned with human-labeled finance-related phrases to classify text sentiment. The study demonstrates that sentiment index derived from such

model can explain abnormal stock returns.

However, the reliance on one-dimensional textual vectors introduces significant information loss. Furthermore, these methods depend on human-labeled datasets and an additional supervised classification step, which amplifies uncertainty. For instance, Malo et al. (2013) highlight discrepancies in human-labeled datasets, underlining the limitations of such approaches. In this study, we show that these methods cannot outperform even simple HAR models.

An alternative approach bypasses dimensionality reduction, avoiding the uncertainty of sentiment estimation, by directly utilizing traditional frequency-based methods with high-dimensional vectors. For instance, Manela and Moreira (2017) applies this technique to back-cast financial market disasters using media news. However, these sparse high-dimensional vectors are constructed using the bag-of-words technique, which fails to account for the semantic and syntactic complexities of textual data. Moreover, the high dimensionality of these vectors impedes regression algorithms from effectively capturing the dependencies between dependent and independent variables.

We propose a textual regression approach that leverages large language models (LLMs) to transform media news text into dense vectors, which are then used to predict multi-horizon-ahead realized volatility of corn, soybean, and wheat. This approach harnesses the state-of-the-art capabilities of LLMs to quantify the information within the text. LLMs address the shortcomings of bag-of-words (BoW)-based high-dimensional vectors by accounting for semantic and syntactic complexities in the text through multiple layers of neural networks stacked to form a structure known as the Transformer. By utilizing this “large” neural network architecture with numerous parameters, LLMs can encode intricate textual information into dense vectors. Furthermore, instead of reducing these vectors into one-dimensional indices, our approach retains all available information in the regression model, effectively circumventing both information loss and additional layers of uncertainty.

The rest of the paper is organized as follows: Section 2 provides a comprehensive methodology of the proposed model, as well as the HAR family of benchmarks. Section 3 reports the empirical analysis, findings, and compares the performance of various models. Section 4 concludes the study. The appendix provides extra explanations, tables, and figures.

2. Methodology

The aim of this study is to investigate whether the information embedded in news articles published by media outlets can forecast the daily realized volatility (RV) of agricultural commodity markets. This can be approached in several ways, such as constructing an index from the text using human-labeled data, including sentiment indices as seen in Araci (2019); Huang et al. (2023); Sinha et al. (2022), and using it as a proxy for the information contained in the text. Alternatively, one could employ news volume as a proxy for media attention on a specific topic, as demonstrated by Engle et al. (2020). However, our approach diverges by regressing daily realized volatility on word vectors derived from the news texts. We argue that by preserving the high-dimensional nature of textual data and avoiding the loss of information typically associated with dimensionality reduction into one-dimensional proxy variables, this method provides superior in-sample fit and out-of-sample forecasting performance.

2.1. Textual regression framework

The proposed textual regression methodology consists of two stages. In the first stage, media news text is transformed into numerical vector representations. In the second stage, these vectors serve as predictors (independent variables) in a regression framework to model and forecast the daily realized volatility of the target agricultural commodities. Formally, the extraction of textual vector representations, $\mathbf{X}$, is conducted as follows:

$$\mathbf{X} = \psi(\text{text}), \quad (1)$$

where ψ represents a text vectorization model, such as LLaMA, BERT-base, BERT-large, RoBERTa-base, RoBERTa-large, FinBERT, FastText, and TF-IDF. A detailed discussion of these natural language processing (NLP) models is provided in Section 2.2.

When processing text data using large language models (LLMs), a key constraint is the context window size—i.e., the maximum number of input tokens¹ the model can handle in a single pass. This limit varies across models. The smallest context sizes belong to models built on the BERT-base architecture, including BERT-base, RoBERTa-base, and FinBERT, which can process up to 512 tokens per input sequence. To accommodate longer news articles, we segment each article into chunks of up to 512 tokens, ensuring that breaks occur at sentence boundaries to preserve semantic integrity. Each chunk is then processed separately, with the LLM generating one vector per token. Finally, to obtain a single vector representation for each news article, we compute the average of the token-level vectors across all chunks within the article. This procedure yields one fixed-dimensional vector per news article, i.e. $\mathbf{X}$ in the above equation, with the same dimensionality as an individual token vector. In general, larger LLMs produce higher-dimensional textual vectors. For instance, LLaMA-2-7B and LLaMA-2-13B generate vectors of 4096 and 5120 dimensions, respectively, while BERT-large and RoBERTa-large yield 1024-dimensional vectors. Meanwhile, BERT-base, RoBERTa-base, and FinBERT produce 768-dimensional vectors².

Since multiple news articles are published each day, an additional challenge arises: determining how to aggregate multiple text-derived predictors for each daily RV observation. Given that the number of daily news articles varies, treating each article as an individual predictor would introduce inconsistencies in the regression framework. To address this, we adopt two approaches. The first method, which we refer to as *textual vector averaging* (TVA), involves computing the mean vector of all news articles published on a given day and using this aggregated vector as the predictor in a regression model. Specifically, we model the relationship between the average textual representation and the h -day-ahead RV as follows:

$$\widehat{RV}_{t+h}^{TVA} = \beta_0 + \beta_1 \frac{1}{n \times p} \sum_{i=1}^n \sum_{j=0}^p \mathbf{X}_{t-j,i}, \quad (2)$$

¹A *token* refers to an individual word or a subword unit used in NLP tasks. Tokenization methods vary across models, with some employing word-level tokenization, while others, such as subword-based models (e.g., BERT and LLaMA), break words into smaller components to handle out-of-vocabulary words more effectively.

²Different terminologies are used to describe these text-derived numerical representations, including embeddings and features. For the sake of consistency and simplicity, we refer to them as vectors throughout this study.

where $\widehat{RV}_{t+h}^{TVA}$ denotes the forecasted realized volatility for day $t + h$ with $h \in \{1, 5, 10, 22\}$ representing one-day, one-week, two-weeks, and one-month-ahead forecasting horizon. Additionally, $X_{t-p,i}$ represents the $m \times 1$ vectorized representation of news article i^3 in day $t - p$, where $p \in N \cap [0, l]$ denotes the lag length, indicating the number of preceding days' news articles incorporated in the regression. The parameter m corresponds to the dimensionality of the textual vectors, which varies across different text vectorization methods, while n represents the total number of news articles published on a given day.

Averaging high-dimensional textual vectors, as in the previous approach, may lead to an excessive smoothing of information, potentially diluting valuable signals. To mitigate this, our second approach predicts the realized volatility of day $t + h$ separately for each news article i and then takes the mean of these individual forecasts as the final predicted realized volatility for day $t + h$. We refer to this method as *forecast averaging* (FA), and it is formulated as follows:

$$\begin{aligned}\widehat{RV}_{t+h} &= \beta_0 + \beta_1 \mathbf{X}_{t-p,i}, \\ \widehat{RV}_{t+h}^{FA} &= \frac{1}{n \times p} \sum_{i=1}^n \sum_{j=0}^p \widehat{RV}_{t+h,i,j}.\end{aligned}\tag{3}$$

We will later demonstrate that this second regression framework facilitates interpretability, allowing us to identify key textual features that contribute to forecasting accuracy. Furthermore, we acknowledge the potential for look-ahead bias in the forecast averaging approach, $\widehat{RV}_{t+h}^{FA}$. To mitigate this issue, we ensure that the in-sample and out-of-sample datasets are separated strictly based on calendar days rather than observation length.

To estimate the model parameters, we employ a linear regression framework with an L2 regularization penalty, commonly known as Ridge regression (Hoerl, 2020):

$$\text{Minimise } \left(\sum_{i=1}^N (RV_i - \beta_0 - \sum_{j=1}^m \beta_j x_{ij})^2 + \alpha \sum_{j=1}^m \beta_j^2 \right),$$

where α is the regularization parameter that controls the degree of shrinkage applied to the regression coefficients. This regularization helps mitigate potential multicollinearity among predictors and prevents overfitting in our high-dimensional feature space. Ridge regression extends the traditional ordinary least squares (OLS) approach by introducing a penalty term that discourages excessively large coefficients, thereby striking a balance between model complexity and predictive accuracy. While nonlinear regression methods, such as random forests, have demonstrated superior predictive performance in various forecasting tasks, they often come at the expense of increased model complexity, higher estimation uncertainty, and reduced interpretability. Given our primary objective—to explore the relationship between textual information and volatility fluctuations—along with the need for result replicability, we prioritize a linear framework over more complex alternatives.

³If a news article exceeds the context size of the language model, i may refer to a chunk of the original text.

We compare our proposed model with Heterogeneous Autoregressive (HAR). First introduced by Corsi (2008), HAR models are a powerful family of autoregressive forecasting models used for realized volatility. The basic form of the HAR model is expressed as:

$$\widehat{RV}_{t+h}^{HAR} = \beta_0 + \beta_1 RV_t + \beta_2 \overline{RV}_t^w + \beta_3 \overline{RV}_t^m + \epsilon_{t+h}, \quad (4)$$

where $\widehat{RV}_{t+h}$ represents the h -day-ahead forecast of realized volatility, RV_t denotes the realized volatility at day t , and $\overline{RV}_t^w$ and $\overline{RV}_t^m$ are the averages of realized volatility over the past week and month, respectively. In addition to this basic form, we utilize Heterogeneous Autoregressive model with Jumps (HAR-J), continuous-HAR (CHAR) (Corsi et al., 2008), semivariance-HAR (SHAR) (Patton and Sheppard, 2015), ARQ (Autoregressive Realized Quarticity), HARQ (Heterogeneous ARQ), and HARQ-F (Bollerslev et al., 2016).

2.2. Large language models

Large language models (LLMs) are deep neural networks built upon the transformer architecture, consisting of multiple layers designed for sequential data processing. Each transformer model is composed of an encoder and a decoder block, but in practice, models often utilize only one of these components. Encoder-only models, such as BERT and RoBERTa, specialize in bidirectional text understanding, while decoder-only models, such as GPT, LLaMA, and OPT, are designed for generative tasks. In this study, we use LLaMA (Touvron et al., 2023) and OPT (Zhang et al., 2022) as the decoder-only models and BERT Devlin et al. (2019), RoBERTa (Liu et al., 2019), and FinBERT (Araci, 2019; Huang et al., 2023) as encoder-only models. When working with LLMs, users have the option to either train a model from scratch or utilize a pre-trained model for downstream tasks. Training an LLM from scratch is computationally expensive and requires vast amounts of textual data, making it impractical for most users. Instead, it is common practice to use pre-trained models, which can be applied directly, further trained to enhance domain-specific expertise, or fine-tuned for specialized tasks.

Further training is a self-supervised process applied to a pre-trained model with general language knowledge. This involves exposing the model to additional domain-specific data to improve its understanding of specialized vocabulary and context. During this phase, the model’s weights are updated through continued optimization. Fine-tuning, in contrast, is a supervised learning approach where additional layers are introduced on top of the transformer architecture to perform tasks such as classification or regression. The original LLM weights remain unchanged or undergo minimal adjustments, while the newly added layers are optimized for the specific task. Due to the high computational and data requirements associated with pre-training and further training—especially for models exceeding several hundred million parameters—LLMs are often used in their original form or after fine-tuning for specific applications.

In addition to the LLMs, we compare our results with traditional NLP models like Term Frequency-Inverse Document Frequency (TF-IDF) based on Bag-of-Words (BOW) and Word2Vec (Mikolov et al., 2018).

Table 1: Descriptive statistics of daily news volume in our sample period, spanning from January 1, 2009, to December 31, 2023.

		Min.	Mean	Median	Max.	Std.
Corn	Full sample	0	51.8604	41	262	45.4069
	In-sample	0	31.0376	29	122	27.7801
	Out-of-sample	0	70.0688	74	262	49.7953
Soybean	Full sample	0	51.5704	40	262	45.5094
	In-sample	0	31.0376	29	122	27.7801
	Out-of-sample	0	69.5251	74	262	50.1668
Wheat	Full sample	0	56.9049	48	262	48.7868
	In-sample	0	31.0376	29	122	27.7801
	Out-of-sample	0	79.5245	87	262	51.8675

3. Empirical analysis

3.1. Data

3.1.1. News data

We collect news articles published on Refinitiv’s Eikon platform under the agricultural category between 2009 and 2023. The initial unfiltered dataset comprises 619,109 articles. Table A.1 summarizes the key attributes of the dataset, while Table A.2 presents the number of articles covering the commodities of interest in this study. Notably, a single news article may be associated with multiple products.

A major challenge in working with textual data is its inherently noisy nature, often containing substantial redundancy and irrelevant information. To refine the dataset, we filter out articles from sources deemed irrelevant to the target commodities. Additionally, we apply regular expressions (Regex) to clean the text. Regex enables pattern-based text processing, allowing us to systematically remove or modify predefined elements within the text. Specifically, we retain only articles containing at least one mention of the target commodities and eliminate email addresses, URLs, dates, boilerplate text, tables, attachments, and any paragraph with fewer than three words. Moreover, we exclude articles that consist solely of headlines without accompanying body text.

Our dataset spans 4,932 trading days from January 2009 to December 2023. Table 1 provides a summary of the number of trading days with news coverage and the total volume of articles published within the sample period. With 5,478 total calendar days in the sample, at least one article was published on 90% of the days, including weekends and bank holidays. On average, approximately 51.5 news articles were published per day.

Lastly, it is important to note that the news datasets for wheat, corn, and soybeans may overlap. A single article can simultaneously appear in all three datasets if it discusses multiple commodities, ensuring comprehensive coverage of market-relevant news.

We partition our full dataset into in-sample and out-of-sample sets based on calendar dates rather than the number of observations. The in-sample dataset comprises the first

Table 3: Descriptive statistics of the full sample of realized volatility for the period from January 1, 2009, to December 31, 2023.

	Min.	Mean	Median	Max.	Std.	Skewness	Kurtosis
Corn	0.3463	2.5824	1.8293	46.5	2.6816	5.3089	49.7817
Soybean	0.0724	1.5881	1.1801	20.8618	1.4345	4.3363	34.0883
Wheat	0.5893	3.5914	2.7022	210.5822	4.7088	25.1143	1029.1034

While these markets share similarities, they also exhibit key differences, particularly in their contract maturity schedules. Corn futures contracts have five delivery months: March, May, July, September, and December. Soybean futures contracts, in contrast, have seven delivery months per year: January, March, May, July, August, September, and November. Wheat futures contracts settle in July, September, December, March, and May. Following Bian et al. (2022b) our analysis for all three commodities focuses on the first nearby futures contract, rolling over to the next available contract as the delivery month begins. Since nearby contracts generally experience higher trading volumes than deferred contracts, they tend to react more swiftly and decisively to new market information.

Consider an asset whose price P_t follows the stochastic differential equation:

$$d\log(P_t) = \mu_t dt + \sigma_t dW_t,$$

where μ_t represents the drift, σ_t is the instantaneous volatility process, and W_t is a standard Brownian motion. The daily integrated volatility from day $t - 1$ to day t is calculated by:

$$IV_t = \int_{t-1}^t \sigma_s^2 ds.$$

This integrated variance is not directly observable. However, it can be approximated using daily realized variance (RV), calculated as:

$$RV_t \equiv \sum_{i=1}^M r_{t,i}^2,$$

where the frequency of sampling $M = 1/\Delta$, and Δ represents the sampling interval. The i -th intraday return for day t is defined as $r_{t,i} \equiv \log(P_{t-1+i\Delta}) - \log(P_{t-1+(i-1)\Delta})$. The RV_t converges to the integrated variance IV_t as $\Delta \rightarrow 0$. However, in practice, increasing the sampling frequency introduces more microstructure noise. The standard practice in the literature is to compute realized volatility based on 5-minute returns (Bollerslev et al., 2016). Justification and discussion on this common choice is provided by Li and Xiu (2016). Table 3 presents descriptive statistics for the RV of the three agricultural commodities under study, and Figure 2 displays the price and volatility dynamics of these crops over the full sample period.

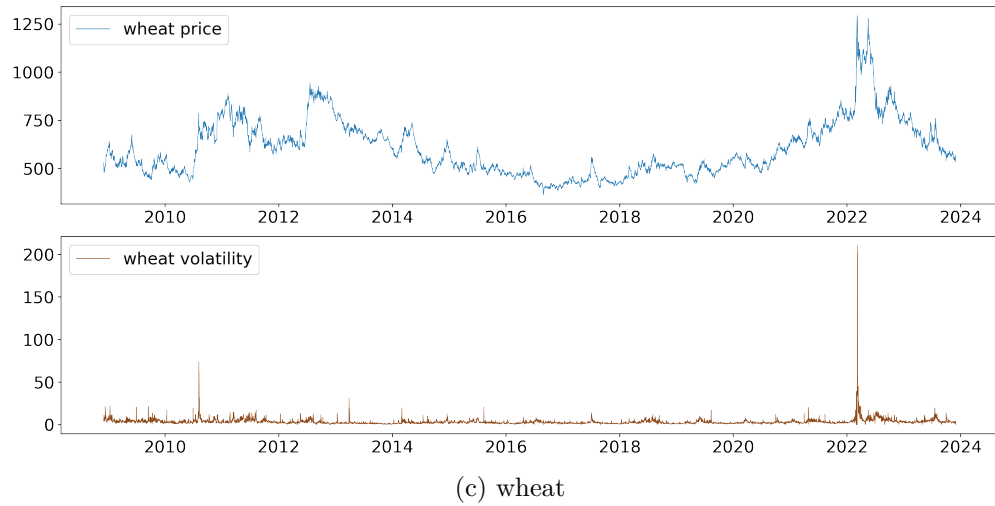
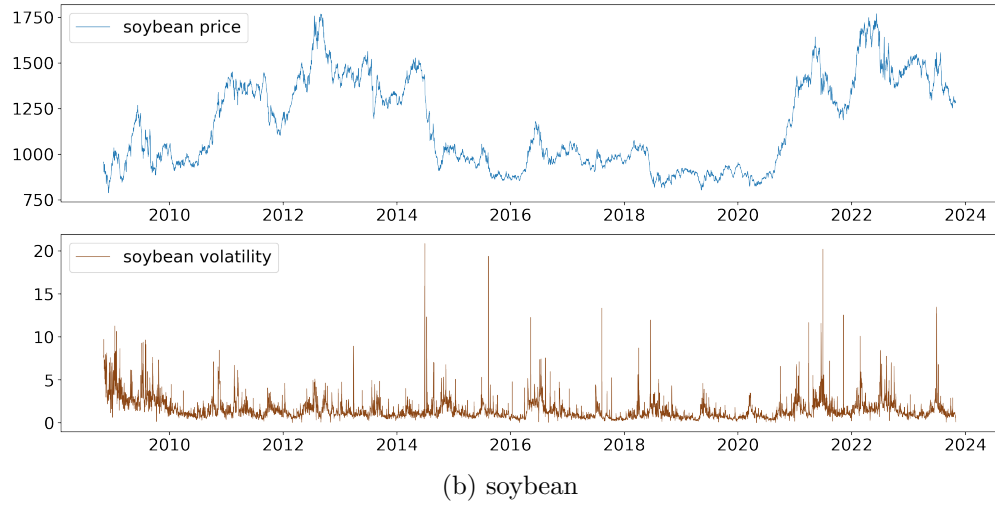
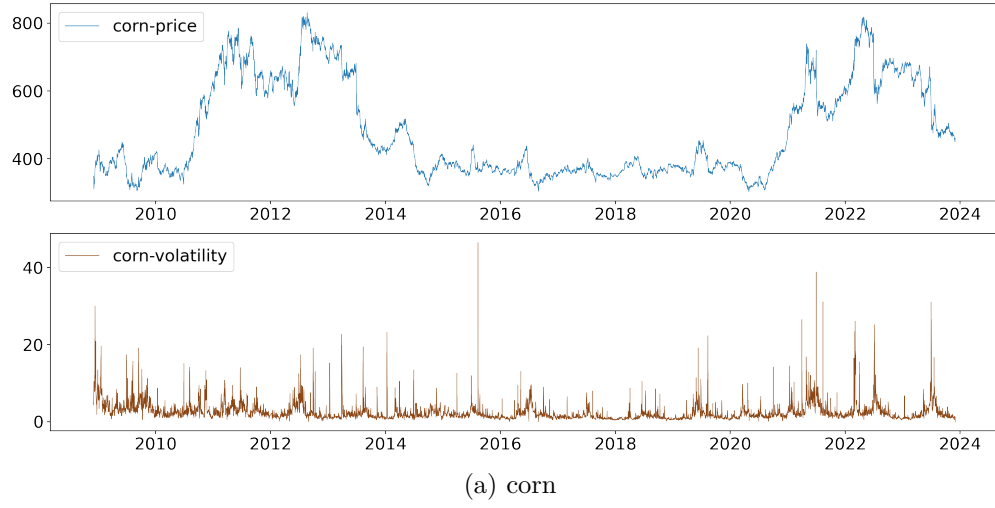


Figure 2: Full sample daily price (upper panels) and volatility (lower panels) for corn (a), soybean (b), and wheat (c) from January 1, 2009, to December 31, 2023.

3.2. Error measures

In line with the OLS model estimations, we employ two metrics—Mean Squared Error (MSE) and Quasi Likelihood (QLIKE)—to evaluate the accuracy of the in-sample and out-of-sample model fits:

$$MSE(RV_t, \widehat{RV}_t) \equiv (RV_t - \widehat{RV}_t)^2,$$

$$QLIKE(RV_t, \widehat{RV}_t) \equiv \frac{RV_t}{\widehat{RV}_t} - \log\left(\frac{RV_t}{\widehat{RV}_t}\right) - 1.$$

For out-of-sample predictions, we compute both MSE and QLIKE, where $\widehat{RV}_t$ represents the forecast from the different models. Patton (2011) explored a range of error measures and found that MSE and QLIKE are the only classes of loss functions that can provide unbiased measurements and correctly rank volatility forecasts in the presence of noise. While MSE treats all errors symmetrically, the QLIKE loss function imposes a greater penalty on large under-forecasting compared to over-forecasting.

In certain instances, the forecast of realized variance $\widehat{RV}$ may become negative, causing the QLIKE loss to be undefined due to the logarithmic function. A reasonable solution to this issue is to replace $\widehat{RV}_t$ with the forecast from the previous observation, $\widehat{RV}_{t-1}$.

For statistical testing of the model outcomes, we use the Model Confidence Set (MCS) approach, as introduced by Hansen et al. (2011). The MCS identifies a subset of models containing the best-performing model with a specified confidence level. Unlike traditional model selection criteria, the MCS adapts to the informativeness of the data, narrowing the set of models with more informative data and expanding it when data provides less distinction. It also accommodates cases where multiple models may perform equally well.

3.3. Parameter optimization

Unlike machine learning methods, linear statistical techniques have relatively few parameters to optimize, which in our case does not drastically affect forecasting accuracy. However, we still optimize the L2 regularization parameter α in Ridge regression, as well as the number of lagged days of news data to include. Specifically, we grid-search α over the space $\alpha \in \left\{10^{\lfloor \frac{K}{2} \rfloor} \times 5^{K \bmod 2} \mid 0 \leq K \leq 8\right\}$, where $K \in \mathbb{N}$, and the number of news lags p over the set $p \in \{1 \leq p < l\}$, where $l \in \{1, 5, 10, 22\}$. The values of l represent one-day, one-week, two-weeks, and one-month histories of news prior to the forecasting day. For example, if $l = 5$, it means we use news data from the past week, spanning from day t to day $t - 5$, to forecast the realized volatility on day $t + 1$ in one-day-ahead forecasting.

We apply the cross-validation method to our in-sample dataset to determine the optimal combination of these parameters. For the optimization process, we use the LLaMA-2-7B language model. Due to the computational cost associated with optimizing these parameters in a high-dimensional predictor setting, we limit the optimization process to the corn dataset and use the resulting parameters across all crops.

The optimization results indicate that the best performance for both one-day- and one-month-ahead forecasting of corn’s realized volatility is achieved when using news data from the past two weeks (i.e., $1 \leq p < 10$) and setting $\alpha = 50$.

Table 4: Out-of-sample analysis of the selected textual regression and HAR family models for one-day-ahead ($h = 1$) forecasting of RV for corn, soybean, and wheat. ***, **, and * denote the best models according to Hansen et al. (2011) MCS at 90%, 80%, and 70% confidence levels, respectively.

$h = 1$	LLaMA-2-7B (TVA)	BERT	FinBERT (Huang)	Fasttext	TD-IDF	HAR	HAR-j	CHAR	SHAR	ARQ	HARQ	HARQ-F
Panel A: Corn												
MSE	5.5288	6.6937	6.5408	8.1793	6.9636	5.1412	5.0765	5.0483	5.1832	5.2912	4.9859	4.9172***
QLIKE	0.1899	0.2563	0.2436	0.4883	0.2797	0.1565	0.1550	0.1518	0.1565	0.1829	0.1530	0.1491***
Panel B: Soybean												
MSE	1.4596	1.7198	1.6796	1.8081	1.7434	1.3560	1.3359	1.3018	1.3703	1.4165	1.3236	1.2892***
QLIKE	0.1777	0.2330	0.2231	0.2527	0.2370	0.1453	0.1432	0.1391	0.1446	0.1680	0.1417	0.1368***
Panel C: Wheat												
MSE	27.9159***	32.1765	31.5513	34.9900	33.8581	30.5864*	30.7148*	30.1664*	30.6848*	30.7531*	31.1212	168.8555
QLIKE	0.1506	0.2027	0.1880	0.2817	0.2363	0.1182***	0.1199	0.1242	0.1187	0.2693	0.1232	0.3347

3.4. Out-of-sample forecasts

In this section, we focus on one-day and one-month-ahead forecasting horizons, while the results for one- and two-week-ahead forecasts can be found in Table B.3 in the appendix. Tables 4 and 5 present the out-of-sample forecasting results for the one-day and one-month-ahead horizons, respectively. Theoretically, unanticipated news affects asset prices through what Engle and Ng (1993) describe as the news-impact curve: unexpected information—such as a surprise crop report—can trigger disproportionately large spikes in intraday volatility. Our LLM-based textual regressions are designed to capture this transmission channel, along with other similar pathways through which news content influences the volatility process. One-step-ahead forecasting is shown in Table 4 where LLaMA-2-7B, although the best-performing model among the textual regression models, exhibits lower accuracy compared to the HAR family models. However, this difference is not statistically significant for wheat, where the MCS identifies LLaMA-2-7B as the optimal model based on the MSE and within the 90% confidence interval. Additionally, HARQ-F outperforms all models in the short-term forecasting of realized volatility for corn and soybean. This result mirrors those found by Bollerslev et al. (2016), who demonstrated the superior forecasting accuracy of HARQ-F for RV in the equity market. This strong performance of HARQ-F is consistent with its underlying jump-augmented framework, which places greater emphasis on large intraday returns—interpreted as sharp market reactions to new information, such as surprise crop reports. In practical terms, volatility in corn and soybean markets tends to exhibit more pronounced spikes around news events, a dynamic that the jump component of HARQ-F is well-suited to capture.

However, when shifting to one-month-ahead forecasting, the results are more nuanced. According to MSE, CHAR emerges as the most accurate model for forecasting corn volatility, a finding consistent with Rahimikia and Poon (2020) in the equity market context. HARQ-F continues to outperform other models in forecasting soybean volatility, while ARQ surpasses HARQ-F in the case of wheat. The inclusion of QLIKE further complicates the interpretation of these results. These mixed outcomes can be attributed to the inherent limitations of autoregressive models in capturing long-term realized volatility. Specifically, it underscores the notion that past volatility observations are insufficient for incorporating information relevant to forecasting more distant future volatility levels.

Over a one-month horizon, the decaying autocorrelation of order flow and the diminishing

Table 5: Out-of-sample analysis of the selected textual regression and HAR family models for one-month-ahead ($h = 22$) forecasting of RV for corn, soybean, and wheat. ***, **, and * denote the best models according to Hansen et al. (2011) MCS at 90%, 80%, and 70% confidence levels, respectively.

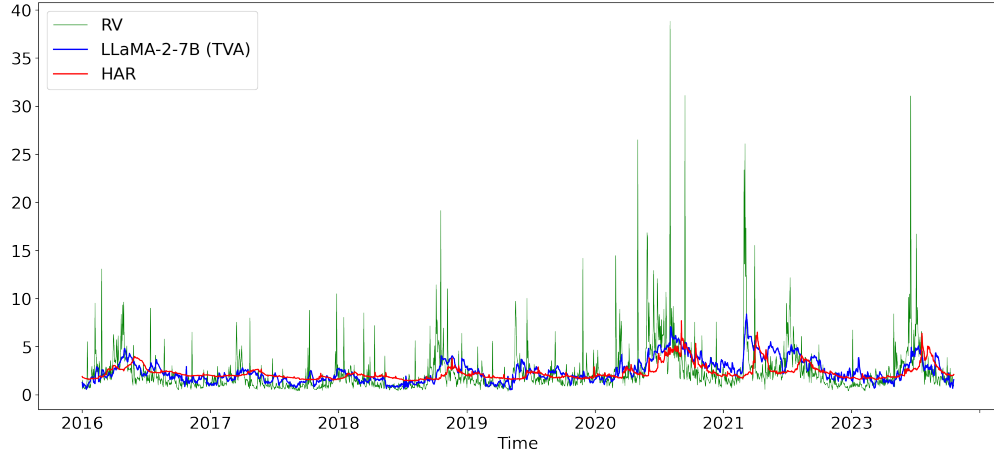
$h = 22$	LLaMA-2-7B (TVA)	BERT	FinBERT (Huang)	Fasttext	TD-IDF	HAR	HAR-j	CHAR	SHAR	ARQ	HARQ	HARQ-F
Panel A: Corn												
MSE	5.6997***	6.7207	6.6017	8.0453	6.8472	6.6927	6.7048	6.6686	6.7031	6.7968	6.6839	6.6819
QLIKE	0.2063***	0.2715	0.2631	0.4907	0.2796	0.2640	0.2700	0.2600	0.2641	0.2727	0.2631	0.2619
Panel B: Soybean												
MSE	1.5243***	1.7453	1.7028	1.8155	1.7688	1.7022	1.7012	1.6835	1.7050	1.7168	1.6966	1.6793
QLIKE	0.1853***	0.2378	0.2269	0.2533	0.2421	0.2262	0.2259	0.2232	0.2264	0.2286	0.2251	0.2223
Panel C: Wheat												
MSE	30.8475***	33.9944	33.5007	35.5934	34.7555	36.4223	36.4048	36.3866	36.7625	34.7082	35.1022	35.5410
QLIKE	0.1741***	0.2230	0.2101	0.2821	0.2458	0.2354	0.2359	0.2368	0.2373	0.2422	0.2333	0.3742

effects of microstructure frictions imply that fundamental news events—precisely what our LLM-based text features are designed to capture—become the dominant drivers of volatility. In the one-month-ahead forecasting results presented in Table 5, textual regression models substantially outperform the autoregressive models. Among the textual models, LLaMA-2-7B emerges as the superior model, outperforming all HAR family models. The MCS test further supports the statistical significance of this superiority across both error measures. LLaMA-2-7B improves long-term volatility forecasting by 17.00% (26.03%), 10.17% (19.97%), and 12.52% (39.12%) for corn, soybean, and wheat, respectively, when comparing its performance to the best-performing HAR family model based on MSE (QLIKE) measures. Furthermore, the QLIKE measure exhibits a steeper increase compared to MSE when transitioning from LLaMA-2-7B to any of the HAR family models, indicating that the HAR models tend to underestimate future volatility—a shortcoming that is penalized more heavily by the asymmetric nature of the QLIKE loss function. Figure 3 illustrates the forecasted values from LLaMA-2-7B and the best HAR family model for each crop during the out-of-sample period.

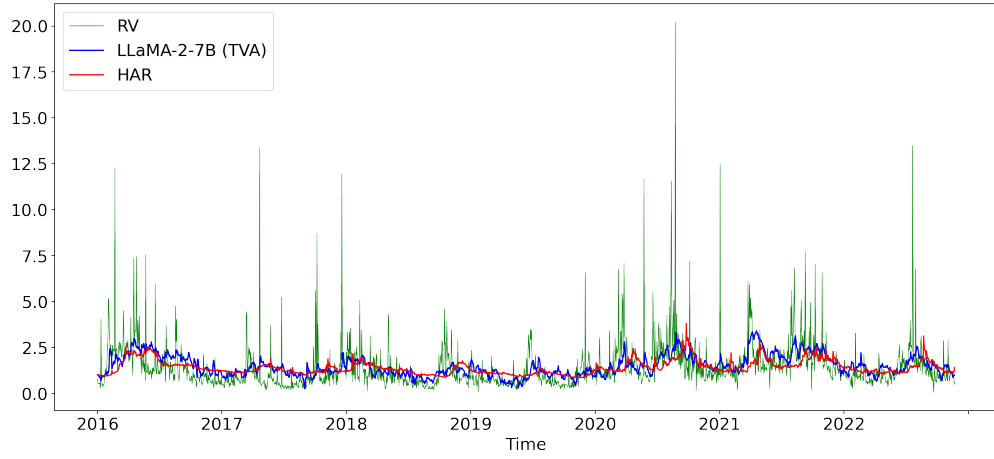
3.5. Exploring other LLMs

A natural question that arises when comparing the forecasting accuracy of LLaMA-2-7B and BERT-base is whether the superior performance of LLaMA-2-7B is due to the differences in the underlying Transformer architecture—the former being encoder-only, while the latter is also encoder-only—or whether it is driven by the substantial difference in model size, with LLaMA-2-7B containing 60 times more parameters. To investigate this, we analyze the impact of various types of language models on multi-horizon forecasting for agricultural commodities. We also explore the effects of LLMs from both encoder-only and decoder-only classes, with varying parameter sizes. Given that the results across the three crops exhibit similar patterns, we focus on presenting the results for corn in this section and the subsequent ones, unless stated otherwise.

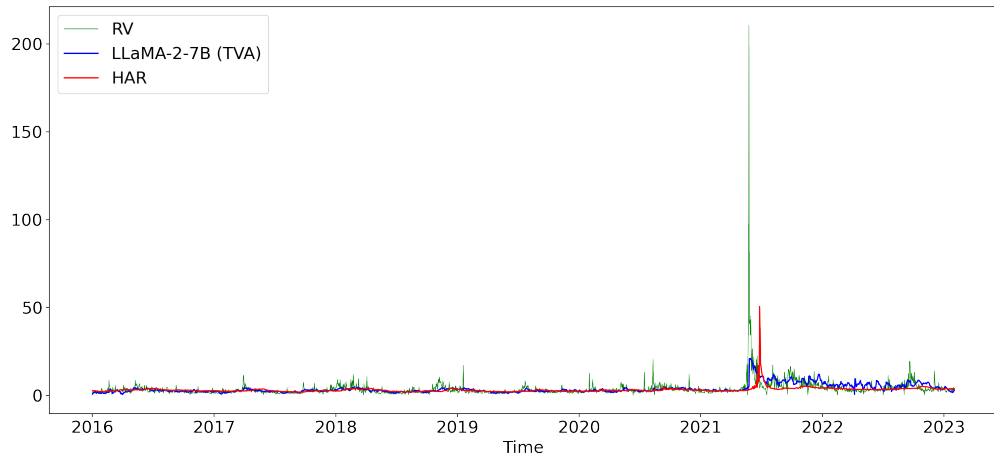
Table 6 reports the results of the out-of-sample regressions for different language models in forecasting the realized volatility of corn. The "FA" and "TVA" refer to the regression frameworks described in Eqs. 3 and 2, respectively. In the FA framework, RV_{t+h} is forecasted by utilizing the text documents from day t . The resulting forecasts for day t , $\widehat{RV}_{t+h,i}$, are



(a) corn



(b) soybean



(c) wheat

Figure 3: One-month-ahead ($h = 22$) out-of-sample forecasting of the LLaMA-2-7B textual regression model (blue lines) versus the best HAR family model (red lines) for corn (a), soybean (b), and wheat (c). The green line represents the observed realized volatility.

Table 6: Comparing the out-of-sample performance of various LLMs in forecasting RV of corn. The "headline" variant refers to a TVA regression setting where the textual vectors are derived from news headlines instead of the full article text. ***, **, and * denote the best models according to Hansen et al. (2011) MCS at 90%, 80%, and 70% confidence levels, respectively.

	Decoder-only								Encoder-only					
	LLaMA-2-7B (TVA)	LLaMA-2-7B (FA)	LLaMA-2-7B (head- lines)	LLaMA-2-13B (TVA)	LLaMA-2-13B (FA)	OPT-125M	OPT-350M	OPT-6.7B	BERT-base	BERT-large	RoBERTa-base	RoBERTa-large	FinBERT (Araci)	FinBERT (Huang)
Panel A: One-day-ahead ($h = 1$)														
MSE	5.5288	5.8364	5.3071***	5.6875	5.7650	6.1581	5.9205	5.6802	6.6937	6.7296	7.1181	6.8257	6.5408	6.2774
QLIKE	0.1899**	0.1982	0.2267	0.1892***	0.1943	0.2173	0.2119	0.1949	0.2563	0.2600	0.2961	0.2672	0.2436	0.2217
Panel B: One-month-ahead ($h = 22$)														
MSE	5.6997	5.8459	5.5737***	5.8542	5.7833	6.2259	6.0084	5.8940	6.7207	6.7210	6.8394	6.7393	6.6017	6.3211
QLIKE	0.2063	0.2046	0.2694	0.2101	0.2005***	0.2345	0.2342	0.2145	0.2715	0.2711	0.2855	0.2722	0.2631	0.2388

subsequently averaged to generate the final forecast $\widehat{RV}_{t+h}$. The TVA framework, on the other hand, involves constructing a single news vector for day t by averaging all the news vectors published on that day, which is then used to forecast RV_{t+h} . Furthermore, the "headline" variant refers to a TVA regression setting in which the textual vectors are derived from the news headlines, rather than the full article text.

Table 6 illustrates that larger models generally yield more accurate predictions, as evidenced by RoBERTa-large outperforming RoBERTa-base, and OPT-7B surpassing OPT-350M, which in turn outperforms OPT-125M. Thus, a clear conclusion emerges that, in general, larger models tend to exhibit higher forecasting accuracy, with the exception of the BERT models. Moreover, the results highlight the efficacy of fine-tuning models for sentiment classification tasks, as seen with Araci’s FinBERT (Araci, 2019) and Huang’s FinBERT (Huang et al., 2023). This additional step optimizes the models for financial text, improving their comprehension of such content, which is reflected in textual vectors of higher quality.

A comparison of LLaMA-2 with BERT and RoBERTa reveals a notable gap between decoder-only and encoder-only models. Given that LLaMA-2 does not have a smaller model comparable to BERT or RoBERTa, and these latter models lack a version of similar size to LLaMA-2, it is crucial to select a model that enables a meaningful comparison. OPT is an appropriate candidate, as it is a decoder-only model with various versions ranging from 125 million to 30 billion parameters. We choose OPT-125m, with 125 million parameters, which is close to the parameter size of BERT-base and RoBERTa-base, OPT-350m (350 million parameters) to compare with BERT-large and RoBERTa-large, and OPT-6.7B (6.7 billion parameters), which is of a similar magnitude to LLaMA-2-7B. Despite both LLaMA-2 and OPT being decoder-only models, they differ in architecture and learning objectives. The results in Table 6 demonstrate that the regression performance of OPT-6.7B closely matches that of LLaMA-2-7B. Therefore, OPT serves as a valid representative of decoder-only LLMs, offering a fair comparison between the two transformer types.

When comparing models with similar parameter sizes, the results show that decoder-only models (such as OPT-125m) outperform their encoder-only counterparts (BERT-base and RoBERTa-base), as indicated by the lower errors observed in OPT-125m compared to both BERT-base and RoBERTa-base. Similarly, OPT-350m outperforms BERT-large and RoBERTa-large in terms of forecasting accuracy, reinforcing the conclusion that decoder-only models tend to be more accurate than encoder-only models in the context of agricultural

commodity volatility forecasting. This superiority is consistent with findings in other natural language processing tasks, such as text generation, where all text generation models are based on decoder-only transformers.

Tables B.4 and B.5 present similar findings for soybean and wheat, respectively. The observed relationships between model size and performance hold for these crops as well. However, the impact of model size on forecasting performance is more pronounced in long-term forecasting, with BERT-large and RoBERTa-large outperforming their smaller counterparts (BERT-base and RoBERTa-base) in one-month-ahead forecasting.

3.6. Sentiment as a predictor

Despite limited studies on using textual vectors for regressing financial market time series, the effect of sentiment measures on financial markets has been extensively studied (Araci, 2019; Huang et al., 2018; Loughran and McDonald, 2011; Malo et al., 2013; Sinha et al., 2022). However, the sentiment classification approach results in a loss of information content within the textual vectors, as it compresses the high-dimensional vectors into a one-dimensional sentiment index. Additionally, this approach introduces an extra layer of uncertainty related to sentiment estimation on top of the inherent uncertainty in the regression model. To assess the effectiveness of sentiment measures in forecasting realized volatility, we propose a sentiment regression model, which is then compared to our LLaMA-2-7B (TVA) textual regression model.

The daily sentiment of the news articles is calculated as the average sentiment of all articles published on a given day:

$$\mathbf{S}_t = \frac{1}{n} \sum_{i=1}^n \phi(\mathbf{X}_{t,i}), \quad (5)$$

where ϕ is a classifier algorithm that categorizes the text as either positive, negative, or neutral. The sentiment vector $\mathbf{S}_t = [S_t^{pos}, S_t^{neg}, S_t^{neut}]$ is then used to forecast the h -step-ahead RV as follows:

$$\widehat{RV}_{t+h}^{Sent} = \beta_0 + \beta_1 S_{t-p,t}^{pos} + \beta_2 S_{t-p,t}^{neg} + \beta_3 S_{t-p,t}^{neut}, \quad (6)$$

where $S_{t-p,t}^{pos}$, $S_{t-p,t}^{neg}$, and $S_{t-p,t}^{neut}$ represent the average probabilities of sentiment (positive, negative, and neutral, respectively) from news published between days $t - p$ and t . The classification models used here are the same as those employed in the textual vectorization process (ψ in Eq. 1), with an added classifier layer on top of the final LLM layer.

In addition to the above setting, we also employ traditional dictionary-based methods for extracting sentiment from the news text. These methods rely on manually constructed word dictionaries categorized into different sentiment classes. Each word in a given text is looked up in the dictionary and assigned a sentiment class. We utilize the well-known Loughran-McDonald dictionary Loughran and McDonald (2011) (LMD) for financial sentiment analysis.

Table 7 presents the results of our attempt to forecast the RV of corn using various sentiment measures. The first column displays the results of our base textual regression model (LLaMA-2-7B) for comparison with the sentiment regression models. Among the sentiment regression models, only the FinBERT models fine-tuned by (Araci, 2019) and (Huang et al.,

Table 7: Comparing sentiment regression models in forecasting RV of corn. The LLaMA-2-7B (TVA) model serves as the baseline textual regression model for comparison against the set of sentiment regression models. ***, **, and * denote the best models according to Hansen et al. (2011) MCS at 90%, 80%, and 70% confidence levels, respectively.

	LLaMA-2-7B (TVA)	LLaMA-2-7B (sentiment regression)	BERT-base	BERT-large	FinBERT (Araci)	FinBERT (Huang)	RoBERTa-base	RoBERTa-large	LMD
Panel A: One-day-ahead ($h = 1$)									
MSE	5.5288***	7.2574	7.3188	7.3220	7.1492	7.0919	7.3183	7.1681	7.3179
QLIKE	0.1899***	0.3086	0.3147	0.3151	0.2980	0.2921	0.3147	0.3020	0.3146
Panel B: One-month-ahead ($h = 22$)									
MSE	5.6997***	7.0553	7.1390	7.1398	7.0020	6.9515	7.1384	7.0100	7.1380
QLIKE	0.2063***	0.2999	0.3085	0.3086	0.2947	0.2898	0.3084	0.2968	0.3084

2023) outperform other sentiment models, albeit only marginally. Notably, the magnitude of model parameters is not directly associated with accuracy in sentiment regression; larger models, such as LLaMA-2-7B (sentiment regression), fail to outperform the smaller, fine-tuned FinBERT models. Moreover, even the most complex and computationally expensive LLMs, such as FinBERT (Huang et al.), offer only marginal improvements over simpler methods like LMD. Overall, sentiment-based models do not effectively forecast RV , largely due to the loss of information and the added uncertainty from the sentiment estimation process. Similar results for soybean and wheat are presented in B.6.

3.7. Randomized test

While rolling window sampling and in-sample statistical significance help mitigate the randomness of model outcomes, we seek to further substantiate the robustness of our results through a randomization test. This test evaluates the null hypothesis that there is no temporal association between the independent textual vectors and the target realized volatility, implying that the two variables are independent. To conduct this test, we disrupt the temporal relationship between the two by randomly shuffling the dependent variable, performing forecasting, and recording the forecasting measures. This process is repeated 1,000 times. The p-value is calculated as the proportion of randomized datasets in which the shuffled estimator performs as well as, or better than, the original data. A small p-value indicates rejection of the null hypothesis, suggesting a genuine temporal dependency between the independent and dependent variables that has been leveraged by the estimator to produce accurate forecasts. A large p-value may either reflect an absence of real dependency between the variables or indicate that the estimator failed to exploit any existing dependency for improved predictions. The results of the randomization test for the crops, using LLaMA-2-7B (TVA), are shown in Table 8, where the p-values for both error measures are statistically significant, leading us to reject the null hypothesis of no temporal dependency. This provides strong evidence in favor of a real temporal relationship between the information in the news texts and the target RV , which our forecasting model effectively exploits. The MSE (true model) and QLIKE (true model) columns in Table 8 present the error measures for the un-

Table 8: Results of the randomization test for crops using 1000 shuffled forecasts for one-day-ahead ($h = 1$) and one-month-ahead ($h = 22$) horizons. The MSE (true model) and QLIKE (true model) columns report the error measures for the true unshuffled model, which uses LLaMA-2-7B (TVA) as its LLM. In each of the MSE (QLIKE) general columns, the Mean (randomized), Median (randomized), and Std. (randomized) sub-columns report the mean, median, and standard deviation of the MSE (QLIKE) for the shuffled models, respectively.

	MSE					QLIKE				
	pvalue	MSE (true model)	Mean (ran- dom- ized)	Median (ran- dom- ized)	Std. (ran- dom- ized)	pvalue	QLIKE (true model)	Mean (ran- dom- ized)	Median (ran- dom- ized)	Std. (ran- dom- ized)
Panel A: One-day-ahead ($h = 1$)										
corn	0.0000	5.5288	7.8955	7.8977	0.1276	0.0000	0.1899	1.6408	0.9307	9.2192
soybean	0.0000	1.4596	2.0531	2.0533	0.0330	0.0000	0.1777	0.7327	0.5391	1.4616
wheat	0.0000	27.9159	34.7670	34.7573	0.3538	0.0000	0.1506	2.2939	0.8605	25.7191
Panel B: One-month-ahead ($h = 22$)										
corn	0.0000	5.6997	7.8756	7.8755	0.1184	0.0000	0.2063	1.4484	0.9187	4.1348
soybean	0.0000	1.5243	2.0566	2.0557	0.0327	0.0000	0.1853	1.2160	0.5400	15.1283
wheat	0.0000	30.8475	34.8042	34.8062	0.3570	0.0000	0.1741	1.4184	0.8602	2.2403

shuffled model, while the MSE (QLIKE) general columns display the Mean (randomized), Median (randomized), and Std. (randomized) sub-columns, reporting the mean, median, and standard deviation of the MSE (QLIKE) across the shuffled models. The results for the one- and two-week-ahead forecasting horizons are shown in Table B.7, which exhibit similar patterns to those observed for shorter and longer forecasting horizons.

3.8. Interpreting the model

Although our LLM-based textual regression model forecasts long-term volatility more accurately than its powerful competitors, passes multiple statistical tests, and benefits from the interpretability of linear regression, the language model itself remains an opaque black box. This limitation arises because, while we can identify vector dimensions—independent variables—with large and statistically significant coefficients, these dimensions cannot be directly linked to specific textual components. Consequently, interpretability remains a challenge. Addressing this issue requires either relying on an interpretable but less powerful TF-IDF approach or employing an explainable AI method, such as Shapley values.

However, because textual regression using LLMs is unprecedented in the finance and economics literature, existing Shapley value software packages are primarily designed for classification tasks and lack direct implementations for textual regression. To circumvent this limitation, we transform our regression framework into a classification task, allowing us to leverage Shapley values to interpret the relationship between textual inputs and volatility forecasts. Specifically, we fit the textual vectors to realized volatility, extract the residuals, and categorize them into "good-fit" and "bad-fit" classes based on a threshold ω . Formally, we compute the in-sample residuals ϵ using the forecast averaging (FA) expression in Eq. 3 as:

$$\epsilon_{t,i} = (\beta_0 + \beta_1 \mathbf{X}_{t-p-1,i}) - RV_t,$$

where $\epsilon_{t,i}$ represents the difference between the model’s predicted and the observed RV_t . We then define the classification variable $C_{t,i}$, a binary index time series that assigns each news item i published at time t to the "good-fit" class if the absolute residual is below the threshold ω , and to the "bad-fit" class otherwise:

$$C_{t,i} = \mathbf{1}_{|\epsilon_{t,i}| < \omega},$$

This transformation enables the application of Shapley values to assess the contribution of different textual features to the classification outcome, thereby providing insights into how specific text contribute to volatility predictions.⁴ To establish the threshold ω for classification, we set it equal to the sample mean of residuals. Consequently, any residual satisfying $-\omega < \epsilon < \omega$ is categorized as a "good fit."

In the next step, we estimate a classification model for $C_{t,i}$ using the textual vectors $\mathbf{X}$ as explanatory variables. Specifically, we adopt a logistic regression framework:

$$P_{t,i}(C_{t,i} = 1 | \mathbf{X}_{t-p-1,i}) = \frac{1}{1 + e^{-\boldsymbol{\lambda} \mathbf{X}_{t-p-1,i}}},$$

where $P_{t,i}$ represents the probability that news item i at time t falls into the "good-fit" category, and $\boldsymbol{\lambda}$ denotes the matrix of logistic regression coefficients.

The fitted probabilities from the logistic regression model vary with changes in the input text, enabling the application of the Shapley values method to quantify the contribution of individual words to classification outcomes. Figure 4 presents the estimated Shapley values for one-day-ahead (Panel A) and one-month-ahead (Panel B) volatility forecasts for corn. Due to the computational intensity of the method, we limit the analysis to corn market news from 2023. The interpretation of Figure 4 is straightforward: words with higher Shapley values contribute positively to the classification of residuals as "good fits," meaning their presence improves the model’s ability to predict realized volatility.

As shown in the figure, short-term (one-day-ahead) forecasts are primarily driven by market-related terms such as *cbot*, *bid*, *high*, *cash*, *future*, *auction*, *rise*, *drop*, *seller*, and *low*, collectively accounting for a total Shapley value of 2374.71, compared to 310.63 for terms related to production and weather. In contrast, for long-term (one-month-ahead) forecasting, while market-related words remain relevant, terms associated with production and weather—such as *corn*, *wheat*, *crop*, *grain*, *soybean*, *grains*, *wheat*, *grain*, *supply*, *weather*, and *agriculture*—become more influential, with a cumulative Shapley value of 3,707.53 compared to 849.51 for market-related terms.

Additionally, B.2 illustrates the most detrimental words—those that contribute to the "bad-fit" classification. As expected, many of these words represent noise or contextually irrelevant terms. Interestingly, some agricultural terms, including *soybean*, *corn*, *wheat*,

⁴We employ the FA setting because, in the TVA framework, each textual vector represents the average of all vectors corresponding to news published on a given day. Consequently, the text associated with each vector comprises the entire body of news from that day, which may exceed the context window of the LLM on certain days. In contrast, the FA setting ensures that each vector corresponds to a text segment that never surpasses the smallest LLM context size, thereby mitigating potential truncation issues and preserving the integrity of textual representations.

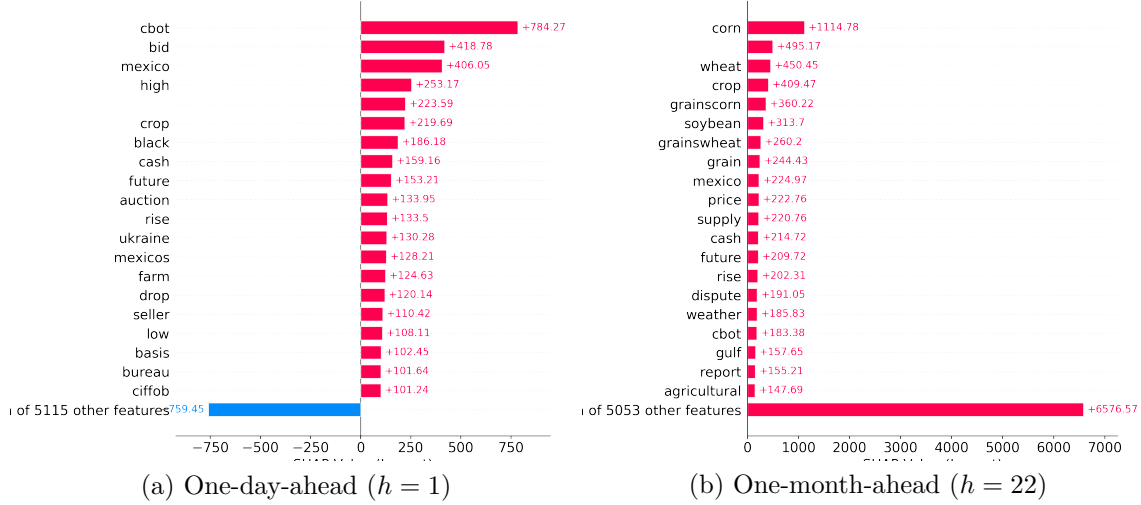


Figure 4: Shapley values representing the most contributing words to the regression of corn volatility in 2023. The figure displays the words with the highest contributions to the model’s forecasting accuracy, based on their Shapley values, highlighting their impact on the realized volatility prediction for corn.

and *grain*, negatively impact short-term forecasting accuracy. This suggests that while fundamental supply-side factors influence long-term volatility, they may introduce noise in short-horizon forecasting, where market microstructure dynamics play a more dominant role.

4. Conclusion

This study investigates the role of textual information in forecasting multi-horizon realized volatility in agricultural markets. To this end, we employ a diverse set of text vectorization methods to transform qualitative news data into numerical representations, which are then used to predict one-day-, one-week-, two-week-, and one-month-ahead realized volatility for corn, soybeans, and wheat. We introduce textual regression models, a novel class of forecasting models that leverage textual information for time-series regression. These models are benchmarked against state-of-the-art autoregressive realized volatility models, particularly the HAR family. To our knowledge, this study is the first to directly utilize textual vectors from LLMs for volatility forecasting across financial markets and the first to explore textual data in agricultural markets.

Our findings reveal that news media is a valuable source of information for long-term volatility forecasting. Textual regression models significantly outperform HAR family models in long-run volatility predictions, as confirmed by both in-sample and out-of-sample analyses. Furthermore, our models surpass sentiment-based approaches—including FinBERT. Hansen’s Model Confidence Set (MCS) test confirms that LLaMA-2 delivers statistically superior performance in long-term forecasting. A randomization test further establishes significant temporal dependence between textual predictors and realized volatility, reinforcing the robustness of our approach.

These findings—that textual regression models consistently outperform autoregressive models, coupled with their relatively stable forecasting accuracy across longer horizons—suggest that markets take time to fully incorporate information conveyed in the news. In con-

trast, the declining performance of autoregressive models as the forecasting horizon lengthens indicates that historical volatility time series lack the necessary information for reliable long-term predictions. In contrast, textual regression models, particularly LLaMA-2-7B, maintain strong performance by directly extracting predictive signals from news text.

Comparative analysis across LLM architectures reveals that forecast accuracy improves with model size. Specifically, decoder-only models, such as OPT, outperform encoder-only models like BERT and RoBERTa of similar sizes in both short- and long-term volatility forecasting. Additionally, interpretability analyses using Shapley values show that short-term forecasts are driven by market-related terms (e.g., *cbot*, *bid*, *high*, *cash*, *future*, *auction*, *rise*, *drop*, *seller*, and *low*), while long-term forecasts rely more on production- and weather-related terms (e.g. *corn*, *wheat*, *crop*, *graincorn*, *soybean*, *grainwhea*, *grain*, *supply*, *weather*, and *agriculture*).

Several caveats qualify the scope and interpretation of our findings. First, our analysis relies on a relatively modest sample of US agricultural news and volatility observations. This limited sample size constrained our ability to explore structural breaks or regime shifts—for example, periods of exceptionally high uncertainty such as extreme weather shocks or trade wars—and prevented us from extending the analysis to other commodity markets or geographical regions. Future work with richer, longer time series could investigate whether the forecasting gains from textual information persist (or even grow) under different market conditions or across global agricultural exchanges. As a result, the external validity of our findings remains an open question, and further research is needed to assess whether large language-model-based text features offer similar forecasting advantages in other settings. Second, we cannot fully rule out endogeneity concerns. In particular, omitted variables—such as macroeconomic announcements, inventory surprises, or unobserved policy interventions—may correlate both with the news vectors and with future volatility, potentially biasing our estimated coefficients. While our primary goal was not to provide the definitive “best” volatility model, but rather to isolate and assess the incremental value of textual features, addressing these endogeneity issues through instrumental variables or richer control sets would strengthen causal interpretation.

References

- Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4):885.
- Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96(453):42–55.
- Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. eprint: 1908.10063.
- Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

- Barndorff-Nielsen, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(2):253–280.
- Berger, J., Dalheimer, B., and Brümmer, B. (2021). Effects of variable eu import levies on corn price volatility. *Food Policy*, 102:102063.
- Bian, S., Serra, T., Garcia, P., and Irwin, S. (2022a). New evidence on market response to public announcements in the presence of microstructure noise. *European Journal of Operational Research*, 298(2):785–800. Publisher: Elsevier BV.
- Bian, S., Serra, T., Garcia, P., and Irwin, S. (2022b). New evidence on market response to public announcements in the presence of microstructure noise. *European Journal of Operational Research*, 298(2):785–800.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.
- Bollerslev, T., Patton, A. J., and Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1):1–18.
- Cao, A. N. Q. and Robe, M. A. (2021). Market uncertainty and sentiment around USDA announcements. *Journal of Futures Markets*, 42(2):250–275. Publisher: Wiley.
- Chen, Y., Kelly, B. T., and Xiu, D. (2022). Expected Returns and Large Language Models. *SSRN Electronic Journal*. Publisher: Elsevier BV.
- Corsi, F. (2008). A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Financial Econometrics*, 7(2):174–196. Publisher: Oxford University Press (OUP).
- Corsi, F., Pirino, D., and Renò, R. (2008). Threshold Bipower Variation and the Impact of Jumps on Volatility Forecasting. *SSRN Electronic Journal*. Publisher: Elsevier BV.
- Dawson, P. J. (2014). Measuring the volatility of wheat futures prices on the liffe. *Journal of Agricultural Economics*, 66(1):20–35.
- Degiannakis, S., Filis, G., Klein, T., and Walther, T. (2022). Forecasting realized volatility of agricultural commodities. *International Journal of Forecasting*, 38(1):74–96.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*: 1810.04805.
- Dutta, A., Uddin, G. S., Sheng, L. W., Park, D., and Zhu, X. (2024). Volatility dynamics of agricultural futures markets under uncertainties. *Energy Economics*, 136:107754.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987.

- Engle, R. F., Giglio, S., Kelly, B., Lee, H., and Stroebe, J. (2020). Hedging climate change news. *The Review of Financial Studies*, 33(3):1184–1216. Publisher: Oxford University Press (OUP).
- Engle, R. F. and Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *The Journal of Finance*, 48(5):1749.
- Gallic, E. and Vermandel, G. (2020). Weather shocks. *European Economic Review*, 124:103409.
- Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.
- Hoerl, R. W. (2020). Ridge Regression: A Historical Context. *Technometrics*, 62(4):420–425. Publisher: Informa UK Limited.
- Huang, A. H., Lehar, R., Zang, A. Y., and Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science*, 64(6):2833–2855. Publisher: Institute for Operations Research and the Management Sciences (INFORMS).
- Huang, A. H., Wang, H., and Yang, Y. (2023). Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841. Publisher: Wiley.
- Indriawan, I., Martinez, V., and Tse, Y. (2021). The impact of the change in USDA announcement release procedures on agricultural commodity futures. *Journal of Commodity Markets*, 23:100149. Publisher: Elsevier BV.
- Karali, B. and Power, G. J. (2013). Short- and long-run determinants of commodity price volatility. *American Journal of Agricultural Economics*, 95(3):724–738.
- Li, J. and Xiu, D. (2016). Generalized Method of Integrated Moments for High-Frequency Data. *Econometrica*, 84(4):1613–1633. Publisher: The Econometric Society.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. [_eprint: 1907.11692](https://arxiv.org/abs/1907.11692).
- Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65. Publisher: Wiley.
- Luo, J., Klein, T., Ji, Q., and Hou, C. (2022). Forecasting realized volatility of agricultural commodity futures with infinite hidden markov har models. *International Journal of Forecasting*, 38(1):51–73.
- Malo, P., Sinha, A., Korhonen, P., Wallenius, J., and Takala, P. (2013). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796. Publisher: Wiley.

- Manela, A. and Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1):137–162. Publisher: Elsevier BV.
- Mikolov, T., Grave, E., Bojanowski, P., Puhersch, C., and Joulin, A. (2018). Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- Mustafa, Z., Vitali, G., Huffaker, R., and Canavari, M. (2023). A systematic review on price volatility in agriculture. *Journal of Economic Surveys*, 38(1):268–294. Publisher: Wiley.
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256. Publisher: Elsevier BV.
- Patton, A. J. and Sheppard, K. (2015). Good Volatility, Bad Volatility: Signed Jumps and The Persistence of Volatility. *Review of Economics and Statistics*, 97(3):683–697. Publisher: MIT Press - Journals.
- Rahimikia, E. and Poon, S.-H. (2020). Alternative data for realised volatility forecasting: Limit order book and news stories. *Available at SSRN*, 3684040.
- Sanjuán-López, A. I. and Dawson, P. J. (2017). Volatility effects of index trading and spillovers on us agricultural futures markets: A multivariate garch approach. *Journal of Agricultural Economics*, 68(3):822–838.
- Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2022). Measuring news sentiment. *Journal of Econometrics*, 228(2):221–243. Publisher: Elsevier BV.
- Sinha, A., Kedas, S., Kumar, R., and Malo, P. (2022). SEntFiN 1.0: Entity-aware sentiment analysis for financial news. *Journal of the Association for Information Science and Technology*, 73(9):1314–1335. Publisher: Wiley.
- Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes—a study of the daily sugar prices 1961–75. *Time series analysis: theory and practice*, 1:203–226.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *__eprint*: 2307.09288.

- Xiang, Y. and Borjigin, S. (2024). High–low volatility spillover network between economic policy uncertainty and commodity futures markets. *Journal of Futures Markets*, 44(8):1295–1319.
- Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. (2022). Opt: Open pre-trained transformer language models.

Appendix A. Data cleansing

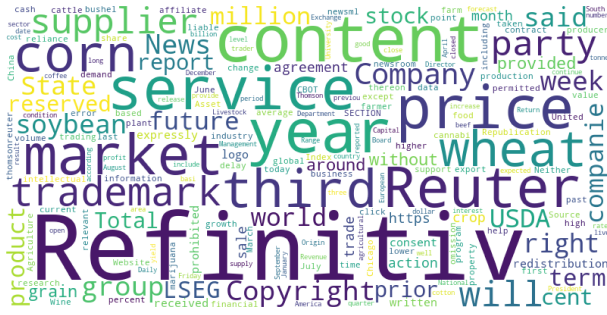
Appendix A.1. Text data statistics

Table A.1: The attributes of the news dataset used in this study. The "News articles" column presents the total number of news articles published each year. The "News without body" column refers to articles that only contain a headline and lack a body of text, often published promptly after significant events. The "News with inconsistent headlines" column identifies articles whose headlines were modified after publication. The "Sources" column indicates the number of distinct news sources or agencies contributing to the dataset.

	News articles	News without body	News with inconsistent headlines	News with sentiment	Sources
2009	19437	4	0	0	1
2010	18030	2	0	0	1
2011	22157	10	10	0	2
2012	11672	7	2	0	2
2013	9812	73	4	0	2
2014	9472	18	0	0	28
2015	11640	17	0	0	39
2016	49790	26	10	0	645
2017	63541	18	8	0	709
2018	39652	3	350	254	454
2019	52739	6	25	0	782
2020	44449	10	14	66	719
2021	58006	562	1135	13729	979
2022	103712	21402	6366	22980	1761
2023	105387	22004	6196	20842	2191

Table A.2: The yearly volume of published news articles and the corresponding word counts before and after the cleansing process.

	Before cleansing		After cleansing	
	Num. news	Num. words	Num. news	Num. words
2009	19433	9612652	16470	531146
2010	18028	8566367	14679	387618
2011	22147	11631355	16451	719524
2012	11665	5805702	8516	392357
2013	9785	4314035	7901	281656
2014	9454	4466632	7513	241884
2015	11619	5615861	7802	168083
2016	49743	40533168	28041	1168356
2017	63510	50180651	29266	1623443
2018	39156	33889587	22556	775514
2019	52242	47423979	22818	1229889
2020	44372	51615969	21090	1605597
2021	56298	55719781	26642	2622152
2022	75991	68868295	10233	2972831
2023	85020	80466841	44165	1046891



(a) before cleansing



(b) after cleansing (corn)



(c) after cleansing (soybean)



(d) after cleansing (wheat)

Figure A.1: Word clouds representing the text dataset before (a) and after (b, c, d) the cleansing process for corn, soybean, and wheat. Panel (a) displays the raw text data with all original words, while panels (b), (c), and (d) show the refined word clouds for each crop after the cleansing steps. The word clouds visually highlight the most frequent terms used in the respective datasets, with the size of each word indicating its relative frequency.

Appendix B. Further results

Appendix B.1. Out-of-sample forecasts

Table B.3: Out-of-sample analysis of the selected textual regression and HAR family models for one-week-ahead ($h = 5$) and two-week-ahead ($h = 10$) forecasting of RV for corn, soybean, and wheat. ***, **, and * denote the best models according to Hansen et al. (2011) MCS at 90%, 80%, and 70% confidence levels, respectively.

		LLaMA-2-7B (TVA)	BERT	FinBERT (Huang)	Fasttext	TD-IDF	HAR	HAR-j	CHAR	SHAR	ARQ	HARQ	HARQ-F
Panel A: one-week-ahead ($h = 5$)													
Corn	MSE	5.6965***	6.8390	6.6694	8.2970	7.0916	6.0406	6.0121	5.9666	6.0731	6.3915	5.9778	5.9007
	QLIKE	0.1980	0.2608	0.2465	0.4916	0.2827	0.1961	0.1922	0.1877*	0.1962	0.2295	0.1921	0.1872***
Soybean	MSE	1.4682	1.7309	1.6878	1.8241	1.7636	1.4826	1.4586	1.4400	1.4924	1.5456	1.4533	1.4354***
	QLIKE	0.1796	0.2337	0.2231	0.2544	0.2396	0.1720	0.1689	0.1670	0.1711	0.1915	0.1672	0.1650***
Wheat	MSE	28.7668***	32.4472	31.8196	35.0347	33.9734	34.2448	34.3752	33.9015	34.0961	82.0502	91.7149	95.7695
	QLIKE	0.1489***	0.2077	0.1932	0.2817	0.2387	0.1526**	0.1541	0.1575	0.1569	0.2499	0.1566	0.2292
Panel B: two-week-ahead ($h = 10$)													
Corn	MSE	5.8986***	6.9783	6.9070	8.3820	7.1679	6.4718	6.4688	6.4048	6.4815	6.8875	6.4637	6.4259
	QLIKE	0.2003***	0.2696	0.2633	0.4976	0.2853	0.2289	0.2280	0.2214	0.2290	0.2582	0.2270	0.2220
Soybean	MSE	1.5050***	1.7381	1.6950	1.8219	1.7673	1.5946	1.5885	1.5652	1.5997	1.6777	1.5856	1.5525
	QLIKE	0.1907	0.2343	0.2235	0.2532	0.2399	0.1956	0.1944	0.1913	0.1956	0.2166	0.1939	0.1885***
Wheat	MSE	30.5519***	33.1375	32.6270*	35.3783	34.4051	34.8206*	34.2602*	34.6325*	34.7702*	38.7988*	41.9170*	44.8964*
	QLIKE	0.1577***	0.2136	0.2014	0.2814	0.2410	0.1850	0.1865	0.1859	0.1859	0.2181	0.1845	0.1812

Table B.4: Comparing the out-of-sample performance of various LLMs in forecasting RV of soybean. The "FA" and "TVA" refer to the regression frameworks described in Eqs. 3 and 2, respectively. In the FA framework, RV_{t+h} is forecasted by utilizing the text documents from day t . The resulting forecasts for day t , $\widehat{RV}_{t+h,i}$, are subsequently averaged to generate the final forecast $\widehat{RV}_{t+h}$. The TVA framework, on the other hand, involves constructing a single news vector for day t by averaging all the news vectors published on that day, which is then used to forecast RV_{t+h} . The "headline" variant refers to a TVA regression setting where the textual vectors are derived from news headlines instead of the full article text. ***, **, and * denote the best models according to Hansen et al. (2011) MCS at 90%, 80%, and 70% confidence levels, respectively.

	Decoder-only							Encoder-only						
	LLaMA-2-7B (TVA)	LLaMA-2-7B (FA)	LLaMA-2-7B (head-lines)	LLaMA-2-13B (TVA)	LLaMA-2-13B (FA)	OPT-125M	OPT-350M	OPT-6.7B	BERT-base	BERT-large	RoBERTa-base	RoBERTa-large	FinBERT (Araci)	FinBERT (Huang)
Panel A: One-day-ahead ($h = 1$)														
MSE	1.4596	1.5108	1.4344***	1.4990	1.4912	1.6085	1.5631	1.4965	1.7198	1.7207	1.7727	1.7235	1.6796	1.5961
QLIKE	0.1777***	0.1858	0.1792**	0.1844	0.1822*	0.2083	0.2051	0.1886	0.2330	0.2319	0.2439	0.2343	0.2231	0.2032
Panel B: One-month-ahead ($h = 22$)														
MSE	1.5243***	1.5487	1.5217***	1.5584	1.5295***	1.6583	1.6355	1.5615	1.7453	1.7396	1.7855	1.7476	1.7028	1.6468
QLIKE	0.1853***	0.1933	0.1974	0.1956	0.1894	0.2172	0.2440	0.1925	0.2378	0.2352	0.2457	0.2385	0.2269	0.2145

Table B.5: Comparing the out-of-sample performance of various LLMs in forecasting RV of wheat. The "FA" and "TVA" refer to the regression frameworks described in Eqs. 3 and 2, respectively. In the FA framework, RV_{t+h} is forecasted by utilizing the text documents from day t . The resulting forecasts for day t , $\widehat{RV}_{t+h,i}$, are subsequently averaged to generate the final forecast $\widehat{RV}_{t+h}$. The TVA framework, on the other hand, involves constructing a single news vector for day t by averaging all the news vectors published on that day, which is then used to forecast RV_{t+h} . The "headline" variant refers to a TVA regression setting where the textual vectors are derived from news headlines instead of the full article text. ***, **, and * denote the best models according to Hansen et al. (2011) MCS at 90%, 80%, and 70% confidence levels, respectively.

	Decoder-only							Encoder-only						
	LLaMA-2-7B (TVA)	LLaMA-2-7B (FA)	LLaMA-2-7B (head-lines)	LLaMA-2-13B (TVA)	LLaMA-2-13B (FA)	OPT-125M	OPT-350M	OPT-6.7B	BERT-base	BERT-large	RoBERTa-base	RoBERTa-large	FinBERT (Araci)	FinBERT (Huang)
Panel A: One-day-ahead ($h = 1$)														
MSE	27.9159	29.1140	27.2362***	28.4760	28.8694	30.1793	29.0313	28.2579	32.1765	32.2716	33.9649	33.1687	31.5513	31.1070
QLIKE	0.1506	0.1484	0.1366***	0.1437	0.1459	0.1641	0.3070	0.1456	0.2027	0.2035	0.2420	0.2185	0.1880	0.1794
Panel B: One-month-ahead ($h = 22$)														
MSE	30.8475	31.9194	29.8905***	31.4494	31.5724	33.1260	31.9646	31.6874	33.9944	33.9687	34.7505	34.2492	33.5007	33.2591
QLIKE	0.1741	0.1693	0.1623***	0.1714	0.1635***	0.1952	0.1899	0.1754	0.2230	0.2222	0.2455	0.2300	0.2101	0.2025

Table B.6: Comparing sentiment regression models in forecasting RV of soybean and wheat. The LLaMA-2-7B (TVA) model serves as the baseline textual regression model for comparison against the set of sentiment regression models. ***, **, and * denote the best models according to Hansen et al. (2011) MCS at 90%, 80%, and 70% confidence levels, respectively.

		LLaMA-2-7B (TVA)	LLaMA-2-7B (sentiment regression)	BERT-base	BERT-large	FinBERT (Araci)	FinBERT (Huang)	RoBERTa-base	RoBERTa-large	LMD
Panel A: One-day-ahead ($h = 1$)										
Soybean	MSE	1.4596***	1.8073	1.8121	1.8096	1.7917	1.7857	1.8100	1.8099	1.8100
	QLIKE	0.1777***	0.2524	0.2538	0.2531	0.2488	0.2473	0.2532	0.2532	0.2532
Wheat	MSE	27.9159***	34.8864	35.0349	35.0373	34.9179	34.8610	35.0406	35.0406	35.0397
	QLIKE	0.1506***	0.2772	0.2840	0.2842	0.2786	0.2761	0.2843	0.2843	0.2842
Panel B: One-month-ahead ($h = 22$)										
Soybean	MSE	1.5243***	1.8126	1.8163	1.8167	1.8027	1.7887	1.8170	1.8169	1.8169
	QLIKE	0.1853***	0.2526	0.2535	0.2536	0.2500	0.2465	0.2536	0.2536	0.2536
Wheat	MSE	30.8475***	35.4735	35.6423	35.6441	35.6035	35.4785	35.6525	35.6525	35.6518
	QLIKE	0.1741***	0.2764	0.2846	0.2846	0.2827	0.2766	0.2851	0.2851	0.2850

Table B.7: Results of the randomization test for crops using 1000 shuffled forecasts for one-week-ahead ($h = 5$) and two-week-ahead ($h = 10$) horizons. The MSE (true model) and QLIKE (true model) columns report the error measures for the true unshuffled model, which uses LLaMA-2-7B (TVA) as its LLM. In each of the MSE (QLIKE) general columns, the Mean (randomized), Median (randomized), and Std. (randomized) sub-columns report the mean, median, and standard deviation of the MSE (QLIKE) for the shuffled models, respectively.

		MSE				QLIKE				
	pvalue	MSE (true model)	Mean (ran- dom- ized)	Med (ran- dom- ized)	Std (ran- dom- ized)		QLIKE (true model)	Mean (ran- dom- ized)	Median (ran- dom- ized)	Std (ran- dom- ized)
Panel A: one-week-ahead ($h = 5$)										
corn	0.0000	5.6965	7.8718	7.8765	0.1247	0.0000	0.1980	1.5894	0.8966	6.9428
soybean	0.0000	1.4682	2.0558	2.0566	0.0337	0.0000	0.1796	0.7038	0.5331	0.8731
wheat	0.0000	28.7668	34.7839	34.7656	0.3463	0.0000	0.1489	1.5742	0.8294	4.3997
Panel B: two-week-ahead ($h = 10$)										
corn	0.0000	5.8986	7.8812	7.8800	0.1258	0.0000	0.2003	1.2980	0.9172	1.8284
soybean	0.0000	1.5050	2.0589	2.0603	0.0338	0.0000	0.1907	0.7630	0.5471	1.8699
wheat	0.0000	30.5519	34.7912	34.7991	0.3650	0.0000	0.1577	1.9513	0.8382	15.6827

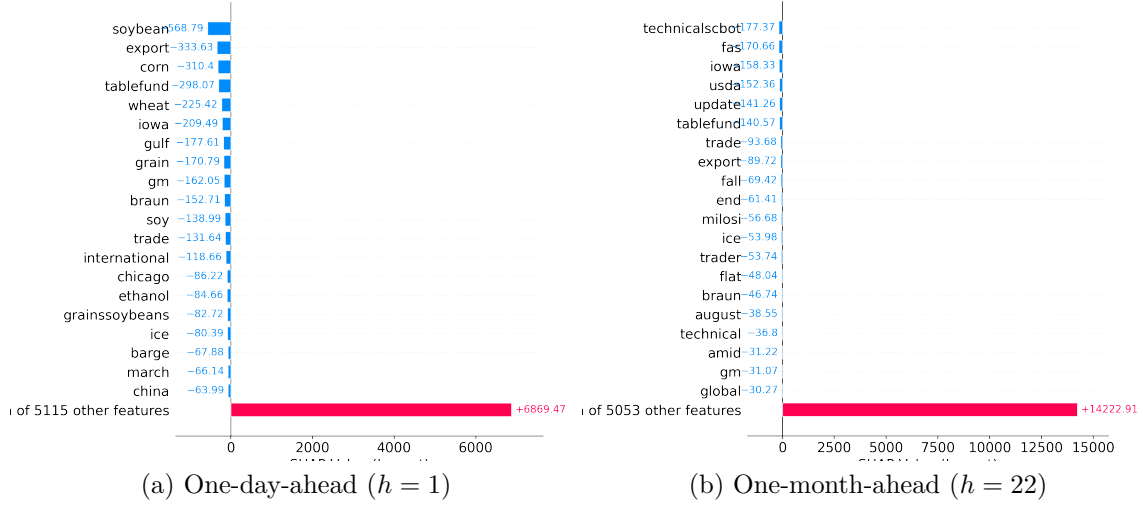


Figure B.2: Shapley values representing the most detrimental words to the regression of corn volatility in 2023. The Shapley values indicate the extent to which each word negatively influences the forecasting model's performance. Words with higher negative Shapley values are those that reduce the accuracy of the volatility predictions when included in the regression, highlighting potential noise or misleading information in the text dataset.