Forecasting Option Returns with News*

Jie Cao, Bing Han, Gang Li, Ruijing Yang, and Xintong (Eunice) Zhan†

July 2024

ABSTRACT

This paper examines the information content of news media for the cross-section of expected equity option returns. Applying various machine learning methods, we derive text-based signals from news articles on publicly traded companies that strongly forecast delta-hedged equity option returns. The option return predictability is robust to variations in methodology and remains significant after controlling for existing predictors. We propose a text-based method to evaluate various underlying mechanisms. We find that media coverage of companies contains valuable information about future change in their stock return volatilities. This appears to be the most important source of option return predictability by news articles.

Keywords: textual analysis, news media, option return predictability, machine learning, large language model

JEL classification: G12, G13, G14, G17

*We thank Turan Bali, Svetlana Bryzgalova, Hector Chan, Hui Chen, Amit Goyal, Yong Jin, Evan Jo, Mete Kilic, Hugues Langlois, Yanchu Liu, Asaf Manela, Dmitriy Muravyev, Neil Pearson, Sang-Ook Shin, Tao Shu, Paul Tetlock, Laurence van Lent, Hua Zhang as well as seminar participants at Chinese Academy of Sciences, Nankai University, Shanghai University of Finance and Economics, University of International Business and Economics, and the Virtual Derivatives Workshop. We have benefited from the comments of participants at CICF (2022), AsianFA (2022), CIRF (2022), SFS Cavalcade Asia-Pacific (2022), FERM (2023), APAD (2023), China Derivatives Youth Forum (2023), Hong Kong Conference for Fintech, AI, and Big Data in Business (2024), 4th Bay Area FinTech Research Forum at Shenzhen University (2024), MFA (2024). The work described in this paper is supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project No. GRF 14500919, 14501720, 14500621, and 15500023) and the National Natural Science Foundation of China (Grant No. 72271061 and 2022HWYQ15). All errors are our own.

†Jie Cao is at Hong Kong Polytechnic University, jie.cao@polyu.edu.hk. Bing Han is at University of Toronto, bing.han@rotman.utoronto.ca. Gang Li and Ruijing Yang are at The Chinese University of Hong Kong, emails: gang.li@cuhk.edu.hk and RuijingYang@link.cuhk.edu.hk. Xintong (Eunice) Zhan is at Fudan University, xintongzhan@fudan.edu.cn.

1 Introduction

Unstructured data, including texts, images, and videos, contain substantial information about firm fundamentals and stock performance. The seminal work of Tetlock (2007, 2010) and Loughran and McDonald (2011) extract information from texts using dictionary-based methods and find that linguistic media contents can capture otherwise hard-to-quantify aspects of firms' fundamentals.[1] Recent studies have delved into employing advanced natural language processing tools in conjunction with machine learning methodologies to obtain insights from unstructured data.[2] Some recent work, such as Ke, Kelly, and Xiu (2019), Kelly, Manela, and Moreira (2021), Frankel, Jennings, and Lee (2022), and Chen, Kelly, and Xiu (2023), highlights that machine-learning methods can yield more potent and reliable asset pricing implications compared to dictionary-based approaches. These studies have focused on the equity market. In comparison, the application of textual analysis in the option market is much limited.[3]

In this paper, we seek to fill the gap by employing machine-learning methodologies to extract information from news media, with the aim of predicting cross-sectional equity option returns based on textual data. Our empirical findings reveal that the information embedded in the news media significantly predicts future equity option returns. In particular, this predictability is distinct from traditional quantitative determinants of option returns and remains robust across various machine-learning algorithms and word feature constructions. Moreover, our study demonstrates the superiority of machine learning approaches in capturing elusive information that is challenging to quantify. This includes information related to implied volatility changes, variance risk premium, implied skewness, and idiosyncratic volatility, which are difficult to be identified using linguistic expressions. Our research also highlights the importance of using alternative data in forecasting equity option returns through the application of machine-learning techniques.

We start our analysis by training a support vector regression (SVR) model following Manela and Moreira (2017) to learn a statistical relationship between news media and future cross-sectional option returns using a massive dataset of enormous news articles on

[1] Tetlock (2007), Tetlock (2010) and Tetlock, Saar-Tsechansky, and Macskassy (2008) show that linguistic media content can capture otherwise hard-to-quantify aspects of firms' fundamentals. Loughran and McDonald (2011) develop a sentiment dictionary that can better reflect the tone of financial text from firms' 10-Ks. Huang, Schlag, Shaliastovich, and Thimme (2019) and Engle, Giglio, Kelly, Lee, and Stroebel (2020) utilize textual analysis to measure firm-level political and climate change risks, respectively.

[2] See Manela and Moreira (2017), Bybee, Kelly, Manela, and Xiu (2021), Bali, Beckmeyer, Mörke, and Weigert (2023), among others.

[3] One notable exception is Manela and Moreira (2017) which use support vector regression and index options data to construct a text-based measure of uncertainty.

stocks with options. SVR is a supervised machine learning algorithm known for its good performance on the ultra-high-dimensional feature space. Applying the trained model out-of-sample, we find that our textual predictors from SVR model can significantly forecast delta-hedged option returns. By sorting options based on the textual signals derived from the SVR model, we find that the average of high-minus-low quintile portfolio return spread is 0.56% (0.36%) per month for call (put) options and robust to controlling for factor models. Moreover, the option-related information extracted from news media is distinct from existing quantitative option return predictors, such as idiosyncratic volatility (Cao and Han (2013)), volatility deviation (Goyal and Saretto (2009)), Amihud illiquidity measure (Amihud (2002)), uncertainty about the implied volatility (Cao, Vasquez, Xiao, and Zhan (2023)), and jump risks proxied by morel-free implied skewness and kurtosis (Bakshi and Kapadia (2003)). The predictive power of textual information for option returns remains robust to various alternative machine-learning methodologies, such as elastic net, random forest, neural networks, and multiple large language models developed recently. In particular, we demonstrate the robustness of our results when applying more advanced text representation techniques, such as word embeddings and large language models suggested by Chen et al. (2023), although the predictive performance is moderate. Our result highlights the usefulness of large language models, but also suggesting a superior performance by the bag-of-words method. In addition to delta-hedged option returns, we also find that textual predictors trained for returns of alternative option portfolios, such as the straddle portfolio and skewness portfolio, exhibit significant predictive power.

We examine potential mechanisms by which news media forecast delta-hedged option returns and propose a novel method to explore what information from news media drives such return predictability. The central idea is to create a dictionary containing important words that help to forecast option returns based on the feature importance through our machine-learning model, and implement a similar process by projecting various option return determinants, such as implied volatility changes, idiosyncratic volatility, volatility deviation, stock illiquidity, jump risks, and uncertainty about the implied volatility, onto the same textual space and construct the corresponding dictionaries for each determinant. Our findings indicate that the textual information contributing to the prediction of option returns stems from a combined effect of multiple sources. Among the projected dictionaries, textual information on future changes in implied volatility represents the most important resource that exhibits approximately 20% lexical overlap with the option-return dictionary, followed by idiosyncratic volatility and implied skewness. Manela and Moreira (2017) demonstrates that news articles can predict changes in implied volatility at the market level. Extending their work, our study shows that news articles are also a valuable source of information for

predicting the dynamics of implied volatility at the individual firm level, and this information significantly contributes to the prediction of equity option returns. We find that a substantial portion of the predictability of news articles on equity option returns is contributed by word features related to idiosyncratic volatility, which primarily captures firm-specific information. This aligns with Cao and Han (2013), which identifies idiosyncratic volatility as a key determinant of option returns. Thus, our findings highlight the importance of firm-specific information from news articles in forecasting equity option returns. Furthermore, our findings indicate that the news contents related to the jump risk significantly contributes to the predictability of the news data on delta-hedged option returns. This result aligns with the conclusions drawn in Jeon, McCurdy, and Zhao (2022), which establishes that material news contents can be important sources of jumps in stock returns.

In addition to our qualitative analysis of word overlap, we utilize quantitative methods to validate and confirm our findings. Specifically, we project each selected option return determinant onto the news media corpus and obtain the fitted value for each variable. We then run regressions of the textual predictors for option returns against these fitted values, quantifying their respective contributions. The results of our quantitative analyzes are consistent with our word-overlap test, reinforcing the robustness and reliability of our findings. The robustness checks further confirm our conclusion that machine-learning approaches excel in extracting information that is challenging to quantify using lexicon-based methods.

Finally, we demonstrate the robust generalizability of our trained SVR model to another important text dataset, earnings conference call transcripts. Specifically, we apply the SVR model trained on news media corpus to earnings conference call transcripts, yielding model-based textual predictors. We find that the model-based textual predictors derived from earnings conference call transcripts exhibit strong predictability for delta-hedged option returns, indicating the reliability and generalizability of our methodologies and analyses.

Our paper contributes to the expansion of the literature on option return predictability, a field predominantly explored in recent studies such as Zhan, Han, Cao, and Tong (2022) and Bali et al. (2023).[4] Diverging from earlier research efforts, we uniquely delve into the realm of textual predictors for option returns, pioneering the analysis of news media alongside the application of machine-learning techniques. Our study demonstrates the robust capabilities of machine-learning in extracting pertinent information from news sources. Methodologically, our research aligns with the approach taken by Bali et al. (2023), who also employ machine learning to forecast option returns, although our distinct contribution lies in our specific focus on textual predictors derived from the news media. This aspect sets our approach apart from

[4]See also Ramachandran and Tayal (2021), Choy and Wei (2023), Jeon, Kan, and Li (2024) for recent developments in this literature

Bali et al. (2023), highlighting the advantages of utilizing machine learning on text data to gain valuable information in predicting future option returns. This distinctive emphasis on textual predictors enriches the existing literature and broadens the understanding of the multifaceted determinants that influence the predictability of the equity option returns.

The remainder of the paper is organized as follows. Section 2 provides sample descriptions and variable constructions. Section 3 provides empirical evidence and robustness checks. Section 4 examines various potential channels and explanations for the option return predictability based on the information derived from news media. Section 5 concludes the paper.

2 Data and Sample

2.1 Data and Sample Description

The newspaper data are mainly collected from ProQuest and complemented with Factiva. At the end of each day, we collect all news articles from the most popular newspapers in the U.S., including the Wall Street Journal, New York Times, Washington Post, and Financial Times. To preprocess the text data, we apply the following steps: First, we filter out any tokens that are not composed of alphabetic characters, such as punctuation marks or numbers. Second, we remove all stop words, which are common words that do not have much meaning, such as "the", "and", or "of". Third, we only keep the words that have a part-of-speech tag of "NOUN", "VERB", "ADJ", or "ADVERB", as these are the most informative and relevant words for our analysis. Fourth, we remove any entities that are recognized by the spaCy module[5], such as names, places, or dates, as they are not useful for our task. By applying these steps, we obtain a clean and concise representation of the text.

Since most articles in ProQuest and Factiva do not have firm-specific tags, we need to identify and match each article with the corresponding firms. We first collect a list of all company names from the Center for Research in Security Prices (CRSP) and conduct a textual fuzzy matching algorithm to search if any firm name appeared (at least twice) in the article.[6] A textual fuzzy match, such as the Jaro-Winkler distance or the Levenshtein distance, is applied to define how similar a specific string is to the target string. We then assign each

[5]spaCy module is a Python package that excels at large-scale information extraction tasks: https://spacy.io/

[6]We perform a placebo test by randomly shuffling the matching of firms and the articles covering them for each month. We then conduct the analysis presented in Table 2. In an untabulated table, we verify that the textual predictors obtained through this process do not exhibit any predictability for future equity option returns

article to its corresponding firms using the textual fuzzy matching algorithm. Note that an article may be assigned to multiple firms since the content may cover multiple companies. To avoid mismatches between news articles and company names, we apply several filters to our data in order to ensure the quality and accuracy of our matching between articles and firms. We exclude those firms that are difficult to identify by company names (e.g., including common words) and remove articles matched with more than seven different companies. We also manually check a random subsample of all company names in our article database and verify that they are correctly matched to their affiliated firms. Before merging with firms engaged in active option trading, we compile a dataset comprising 2,779,518 unique articles, resulting in a substantial article-month sample consisting of 4,462,399 observations.

We obtain equity option data from the OptionMetrics database, which includes information on best bid, best offer, expiration date, and strike price.[7] We also collect variables on underlying stocks, such as stock return, stock price, trading volume, and shares outstanding, from the CRSP database. Our sample period spans from January 1996 to November 2022. In each month, we keep equity options with more than one month until expiration and standard expiration dates. We follow the literature and apply several filters to ensure the quality of our option data. In particular, we exclude observations that breach no-arbitrage limits, have no trading activity in the month preceding portfolio formation, or have zero open interest. Additionally, we discard options that have a mid price lower than \$0.125, have a bid-ask spread lower than the minimum tick size, or involve dividend payment during the holding period (we require that the announcement date of the dividend is no later than the portfolio formation date to avoid any look-ahead bias).[8] Finally, we retain only those options with a moneyness ranging from 0.8 to 1.2.[9] All filters are strictly based on information prior to the portfolio formation date, so no future information is involved in our filtering process. For each firm, we choose options from our filtered set that are closest to being at-the-money. We also ensure that the firms included in our sample have both call and put options available after filtering. The holding period is from the beginning to the end of each month. After merging the newspaper database and the option sample, our final sample consists of 1,061,737 article-month observations and 872,405 unique articles. This dataset is comprehensive and covers a wide range of news media sources for U.S. firms, allowing us to capture the effects of news media contents on option returns more effectively than previous studies.

The final sample contains 89,895 option-month observations for both call and put options

[7] Stock prices are adjusted for stock splits.

[8] \$0.10 for options trading above \$3 and \$0.05 otherwise

[9] Moneyness is defined as the ratio of the strike price (K) to the stock price (S), represented as K/S.

on individual stocks over a 323-month sample period from January 1996 to November 2022. On average, we have 278 option observations for each month. Since we require a firm to be both media-covered and have valid options, our sample consists of mostly large firms. Although our sample contains only 3.69% of the total number of firms in the market, the total market cap of these firms represents 33.55% of the total market. In the universe of optionable stocks, our sample comprises 9.54% of the total number and 36.10% of the total market capitalization of optionable stocks, on average. As shown in Table A1, the firms in our sample rank in the 87^{th} percentile on average in the CRSP stock universe with an average firm size of 33.16 billion. 65.03% of their market shares are held by institutions, and 12.39 analysts on average follow them. In terms of industry composition, our sample is also representative of the whole market.

2.2 Variable Constructions

The main outcome variable of our study is the returns of delta-hedged option portfolios with daily rebalancing, which is the dollar gains of daily-rebalanced delta-hedged long option positions scaled by the initial costs.[10] The delta-hedged call option gain is defined as the change in the value of a self-financing portfolio consisting of a long call position, hedged by a short position in the underlying stock so that the portfolio is not sensitive to stock price movement, with the net investment earning risk-free rate.[11] Specifically, consider a call option that is hedged discretely N times at t_n, $n = 0, 1, \cdots, N-1$ over a period $t, t+\tau$ (where we define $t_0 = t$, $t_N = t+\tau$), its delta-hedged gain is given by

$$\Pi_{t,t+\tau} = C_{t+\tau} - C_t - \sum_{n=0}^{N-1} \Delta_{c,t_n}(S_{t_{n+1}} - S_{t_n}) - \sum_{n=0}^{N-1} \frac{a_n r_{t_n}}{365}(C_{t_n} - \Delta_{c,t_n} S_{t_n}), \tag{1}$$

where C_t is the price of the call option on t, Δ_{c,t_n} is the delta of the call option on t_n, r_{t_n} is the annualized risk-free rate on t_n, and a_n is the number of calendar days between t_n and t_{n+1}. The daily rebalanced delta-hedged put option gain is defined similarly. The delta-hedged option gain $\Pi_{t,t+\tau}$ is the excess dollar return of the delta-hedged option. To make delta-hedged option gains comparable across the underlying stocks, we scale delta-hedged option gains by the initial costs, specifically, $\Delta_{c,t} S_t - C_t$ for call options and $P_t - \Delta_{p,t} S_t$ for

[10] Tian and Wu (2023) show that the daily-rebalanced delta-hedging strategy can remove as much as 90% of the return variation of naked option portfolios. Several previous papers study the delta-hedged option returns, such as Cao and Han (2013), Ramachandran and Tayal (2021), Zhan et al. (2022), and Bali et al. (2023).

[11] If the delta for an option is missing from the OptionMetrics data on a given day, we use the current stock price and the most recent non-missing implied volatility to estimate the delta.

puts. Hence, our delta-hedged option returns are defined as:

$$
\begin{aligned}
r_{i,t}^{\text{call}} &= \frac{\Pi_{t,t+\tau}}{\Delta_{c,t} S_t - C_t} \\
r_{i,t}^{\text{put}} &= \frac{\Pi_{t,t+\tau}}{P_t - \Delta_{p,t} S_t}.
\end{aligned}
\tag{2}
$$

We use text data from news articles to predict the delta-hedged option returns. Following Manela and Moreira (2017), we first build an extensive set of potential information unigrams and then, to mitigate the large dimensionality of the data, we adopt the practice outlined by Kelly et al. (2021) by selecting the top 10,000 unigrams based on their frequencies. Since we train our model on a rolling basis, we construct the unigram space independently for each training process rather than using the entire corpus. This approach ensures that our textual predictors are devoid of future information, relying solely on data available at the training stage. Table A2 demonstrates that the top 10,000 unigrams already represent more than 96% of the unigram frequency.

We follow Kelly et al. (2021) and merge all the news articles covering a firm during a given month into a single document. One can construct a simple counting matrix in each month where, for a firm i and a word j, the corresponding entry is the number of counts that the word j appears in news articles about firm i in that month. Instead of using the simple counting matrix, we follow a common practice in the literature, we use the adjusted word count by term frequency-inverse document frequency (tf–idf) for each word j and firm i at time t given by:

$$
w_{i,t}^{j,tfidf} = \begin{cases} 1 + \log(tf_{i,t}^{j}) w_t^{j,idf}, & if \;\; tf_{i,t}^{j} > 0 \\ 0 & otherwise \end{cases} . \tag{3}
$$

Here $tf_{i,t}^{j}$ is the frequency of occurrence of the word j in the news coverage about firm i at time t, and $w_t^{j,idf} = \log \frac{H_t}{df_t^j}$ where $H_t = \sum_{i=1}^{N_t} H_{i,t}$ is the total number of news articles in the sample at time t, and df_t^j is the number of documents in which the word j appears. We use the tf–idf matrix as the input to forecast delta-hedged option returns.

Next, we apply machine-learning (ML) models to the text data. Because we use high-dimensional data, traditional statistical methods (e.g. OLS) do not work well. In a seminal paper, Manela and Moreira (2017) apply the support vector regression to construct a news-based VIX through high-dimensional textual information. Following the technique used by Manela and Moreira (2017), we consider the following linear regression problem in cross-

section at the end of each month:

$$r_{i,t} = \alpha_t + \beta_t' x_{i,t-1} + \epsilon_{i,t}, \qquad i = 1, 2, \ldots, N_t, \tag{4}$$

where $r_{i,t}$ is delta-hedged option returns for firm i over the month $[t-1, t]$, and $x_{i,t-1} = [x^1_{i,t-1}, \ldots, x^K_{i,t-1}]'$ is a $K \times 1$ vector of (all the) K word features from newspaper articles related to firm i at $t-1$. The support vector regression (SVR) can be formulated as follows:

$$\beta_t^* = \arg\min_w \tfrac{1}{2}||\beta_t||_2 + C\sum_{i=1}^{N_t}(\xi_{i,t} + \xi^*_{i,t}),$$

$$subject\ to \begin{cases} r_{i,t} - \beta_t' x_{i,t-1} - \alpha_t \leq \epsilon + \xi_{i,t} \\ \beta_t' x_{i,t-1} + \alpha_t - r_{i,t} \leq \epsilon + \xi^*_{i,t} \quad , i = 1, 2, \ldots, N_t. \\ \xi_{i,t}, \xi^*_{i,t} \quad \geq 0 \end{cases} \tag{5}$$

The assumption is that such a linear function between $r_{i,t}$ and $\beta_t' x_{i,t-1}$ exists and approximates all pairs $(x_{i,t-1}, r_{i,t})$ with ϵ precision. However, optimization is not always feasible because some of the points fall outside the ϵ margin. As such, we need to account for the possibility of errors larger than ϵ. Following Cortes and Vapnik (1995), we introduce slack variables $\xi_{i,t}$, $\xi^*_{i,t}$ to cope with the otherwise infeasible constraints of the optimization problem (i.e., soft margin). The soft margin gives flexibility to define how much error is acceptable to fall outside of ϵ. The constant $C > 0$ determines the trade-off between the flatness of the linear function and the amount up to which deviations greater than ϵ are tolerated. This corresponds to dealing with the so-called ϵ- insensitive loss function $|\xi|_\epsilon$ described by:

$$|\xi|_\epsilon := \begin{cases} 0, & if\ \ |\xi| \leq \epsilon \\ |\xi| - \epsilon, & otherwise \end{cases} . \tag{6}$$

The above problem can be solved in its dual form (see Schölkopf and Smola (2002)). We train our model on a rolling basis to obtain out-of-sample signals. Specifically, we utilize a training sample that spans one year, followed by a validation sample of six months and a test sample of six months. The training set includes all observations from the previous year to fit the SVR model. The subsequent six-month data are reserved for the validation phase, facilitating the fine-tuning of hyperparameters. Afterwards, the fitted model is employed to forecast delta-hedged option returns for the next six months in a cross-sectional manner, constituting the test sample. This process iterates every six months, effectively rolling forward the training, validation, and test samples, and is repeated until November 2022. The fitted value derived from the SVR model in the test sample is considered as the

textual information extracted from the news media regarding future equity option returns. Specifically, we define the textual predictors (TP) for a given firm, denoted as i, regarding its future equity option returns at time $t+1$, based on news media available at time t, as follows:

$$TP_{i,t} \equiv \hat{r}_{i,t+1} = \hat{\alpha}_t + \hat{\beta}_t' x_{i,t}, \qquad i = 1, 2, \ldots, N_t, \tag{7}$$

where $\hat{\alpha}_t$ and $\hat{\beta}_t$ are fitted parameters in Equation (4) based on the SVR model using the training and validation sample. We perform various tests to assess the predictive power of textual predictors for delta-hedged equity option returns. When constructing textual predictors, we train the models separately for delta-hedged call and put option returns. Although the delta-hedged call and put option returns are highly correlated due to the put-call parity relationship, the main drivers can still be different between them.

[Insert Table 1]

Panel B of Table 1 reports the time-series average of the cross-sectional correlations among the SVR textual predictors and many existing option return determinants. Although the SVR textual predictors for calls and for puts have a fairly high correlation (0.78), their correlations with quantitative option return determinants are still low, in general.

In addition to SVR, we also consider other machine learning methods, such as elastic net, random forest, and neural network, to deal with high-dimensional data from the news media and capture potential nonlinearity and interactions among independent variables. We choose SVR as the main machine learning method for our empirical results because it has fewer hyperparameters to tune, making it more interpretable and stable, and less prone to data snooping issues. In Section 3.2.1, we apply alternative machine learning approaches as robustness checks on our empirical results and find that our findings consistently hold across different machine learning models.

3 Empirical Results

3.1 Baseline Results

3.1.1 Single Portfolio Sorting

We employ Support Vector Regression (SVR) to predict equity option returns, leveraging textual information extracted from newspaper articles. The forecasted delta-hedged call and

put option returns, denoted as TP^{Call} and TP^{Put}, respectively, serve as textual predictors. Portfolios are created by dividing firms into quintiles based on their textual predictors. We then evaluate and compare the realized returns of these portfolios in the following month.

[Insert Table 2]

Table 2 shows that the monthly long-short option strategy based on textual predictors generates positive returns that are significant both economically and statistically. For example, the average monthly return spread between the top and bottom quintiles sorted by the textual predictors is 0.56% (0.36%) for call (put) options. This return spread is substantial, comparable to the magnitude of the median monthly returns of daily-rebalanced delta-hedged options, which is 0.29% (0.42%) for call (put) options. We also adjust the portfolio return spreads using two risk factor models. The first model is the seven-factor model used in Boulatov, Eisdorfer, Goyal, and Zhdanov (2022), which includes the five stock factors in Fama and French (2015), the momentum factor, and the option factor in Coval and Shumway (2001). The other model is the two-factor option model from Zhan et al. (2022) which consists of an idiosyncratic volatility factor and an illiquidity factor. Our results are robust to these risk adjustments, and the adjusted alphas match the raw mean returns closely, indicating that common risk factors do not drive our results.

As shown in Table 2, textual information significantly predicts delta-hedged option returns for at least one month. In an untabulated table, we replicate the results of Ke et al. (2019) and show that our text data significantly predict future stock returns at the daily frequency, but the predictive power diminishes rapidly as the horizon increases. Consistent with previous studies, our text data fail to predict stock returns at the monthly frequency. This suggests that the option market assimilates information from news articles more slowly than the stock market.

3.1.2 Double Portfolio Sorting

To examine whether the effects of our SVR textual predictors are robust to controlling for other option return predictors, we extend our portfolio analysis by double sorting options by various option or stock characteristics first and our textual predictors. We consider six control variables, including (1) idiosyncratic volatility (IVOL) estimated from the Fama-French 3-factor model as in Ang, Hodrick, Xing, and Zhang (2006); (2) volatility deviation (HV-IV), computed as the difference between realized volatility and implied volatility of the at-the-money options as in Goyal and Saretto (2009); (3) Amihud stock illiquidity measure (ILLIQ); (4) Model-free implied skewness (MFIS), calculated as in Bakshi and Kapadia

(2003);(5) Model-free implied kurtosis (MFIK), calculated as in Bakshi and Kapadia (2003); (6) Uncertainty about the implied volatility (VOIV), calculated as the standard deviation of the daily percentage change of option implied volatility during the month.

At the end of each month, we first sort all options into tertiles based on one of the control variables. Within each group, we further sort the options into quintiles based on our SVR textual predictors (i.e., TP^{Call} or TP^{Put}). Finally, we average returns for each textual predictor quintile across the groups of control variables, yielding five control-variable adjusted quintile returns. Table 3 shows that none of the above control variables can subsume the effects of our SVR textual predictors. The return spreads of call options range from 0.47% to 0.55% per month, and those of put options range from 0.27% to 0.36% per month, after controlling for each variable separately. These results are statistically and economically significant and confirm the robustness of our main findings.

[Insert Table 3]

Table 3 confirms that the predictive power of news-based option predictors on equity option returns cannot be explained by existing option predictors. In most of the bins sorted by the control variables, we continue to see the average call or put delta-hedged option returns to be positively related to the news-based option predictors, and most of the portfolio spreads between the high and low textual predictor quintiles are statistically significant. The results in Table 3 are robust if we adjust the raw delta-hedged option returns to alphas based on the seven-factor model used in Boulatov et al. (2022) or the option two-factor model in Zhan et al. (2022).

3.1.3 Fama-Macbeth Regression

To further validate the effectiveness of SVR textual predictors derived from the news media in forecasting the cross-sectional option returns, we conduct the Fama and MacBeth (1973) regression to test whether the predictive power of textual predictors for delta-hedged option return is statistically significant, especially after simultaneously controlling for existing option return predictors. For each dependent variable (delta-hedged call or put option returns), we run the following cross-sectional regressions where the key independent variable of interest is the SVR textual predictor:

$$r_{i,t} = \alpha_t + \beta_t TP_{i,t-1} + \sum_{j=1}^{M} \gamma_t^j X_{i,t-1}^j + \epsilon_{i,t}, \quad i = 1, \ldots, N_t, \tag{8}$$

where $r_{i,t}$ is either delta-hedged call or put option returns for firm i at time t. $TP_{i,t-1}$ is the textual predictor (i.e., $\hat{r}_{i,t}$) for firm i at time $t-1$, and $X^j_{i,t-1}$ are control variables that we use to perform double portfolio sorting in Section 3.1.2. All independent variables are winsorized at the 1^{st} and 99^{th} percentiles and standardized cross-sectionally with zero mean and unit standard deviation.

We run the cross-sectional regression of Equation (8) each month. After obtaining the time series of the coefficients (e.g., β_t) for the independent variables, we perform the t-test for each coefficient using Newey and West (1987) standard errors with the four-lag correction. The hypothesis of the t-test is: $H_0 : \beta = 0$ *vs.* $H_a : \beta \neq 0$. The average of the time-series coefficients and the corresponding t-statistics are reported in Table 4.

[Insert Table 4]

The results of Table 4 support our claim that SVR textual predictors contain useful information about future equity option returns, and their predictive power is robust to various controls. In the univariate regression, the coefficient on TP^{Call} (TP^{Put}) is 0.18 (0.12) with t-statistics of 6.35 (5.24). In the multivariate regression, the coefficients on the SVR textual predictors remain economically and statistically significant, with a coefficient on TP^{Call} (TP^{Put}) of 0.12 (0.08) and a t-statistic of 4.50 (3.41). The other coefficients in Table 4 are in line with the existing literature on option return predictability. For instance, idiosyncratic volatility has a negative effect on delta-hedged option returns, while volatility deviation has a positive effect on forecasting the cross-section of equity option returns. In Table A3, we incorporate additional controls for dictionary-based sentiment and uncertainty measures derived from the Loughran and McDonald (2011) dictionary. Table A3 shows that the inclusion of dictionary-based measures can hardly affect the coefficients of our SVR textual predictors.

3.2 Robustness Checks

3.2.1 Alternative Machine-Learning Approaches

So far, we have demonstrated the usefulness of using the news media to forecast equity option returns. However, one potential concern using machine learning is the possibility of overfitting and data mining due to the choice of multiple hyperparameters. To address this concern, we check the robustness of our empirical results with respect to different values of the hyperparameters. For our main results based on Support Vector Regression (SVR), there are two primary tuning hyperparameters: the regularization parameter (C) and the

epsilon (ϵ). C penalizes each misclassified data point. A low C implies a low penalty and a large margin, but more misclassifications. C reflects the regularization strength, which can be an L_2 penalty. ϵ defines the tube around the actual value where no penalty is applied in the loss function. In our main empirical results, we use Optuna, a state-of-the-art hyperparameter optimization framework in Python, to tune the hyperparameters with 100 trials.[12] We conduct robustness checks using different parameters of C and ϵ to train the SVR model. We show in an untabulated table that the predictive power of textual predictors is robust and significant across reasonable values of C and ϵ.

We also check the robustness of our empirical results by varying the number of input variables for the machine-learning model. For example, in constructing the $tf-idf$ matrix, we experiment with various numbers of maximum features, such as 8,000, 6,000, or 4,000 words, diverging from the 10,000 most frequent words used in our main analysis. The results in Table A4 indicate that our findings remain consistent when varying the number of words in the $tf-idf$ matrix. Another variable we adjust is the length of the rolling window for the model. To evaluate the impact of different rolling window lengths on our results, we also test rolling windows of three, nine, and twelve months, and then re-run the SVR to obtain the textual predictors. The results, as shown in Table A5, remain consistent and significant, underscoring the robustness of our findings to variations in the duration of the rolling window.

In addition, to verify that our results are not driven by the specific choice of the machine-learning approach (i.e., SVR), we also apply alternative machine-learning methods such as elastic net, random forest, and neural networks to extract useful information from the news media to predict option returns.

A model choice close to SVR is the elastic net, which has been successfully applied to solve various topics in asset pricing (see, e.g., Chinco, Clark-Joseph, and Ye (2019) and Dong, Li, Rapach, and Zhou (2021)). The model can be expressed as follows:

$$\alpha_t, \beta_t = \underset{\alpha_t \in R, \beta_t \in R^K}{\arg\min} \left\{ \frac{1}{N_t} \sum_{i=1}^{N_t} \left(r_{i,t} - \alpha_t - \sum_{k=1}^{K} \beta_t^k x_{i,t-1}^k \right)^2 + \lambda \sum_{k=1}^{K} |\beta_t^k| + (1-\lambda) \sum_{k=1}^{K} (\beta_t^k)^2 \right\}, \quad (9)$$

where $r_{i,t}$ is the target variable (delta-hedged equity option returns), N_t is the number of firms i in month t, K is the number of word features $x_{i,t-1}^k$ in the news articles, and λ is a hyperparameter that specifies the weights between L_1 norm and L_2 norm in the loss function. The main difference between SVR and the elastic net is that while the loss function of the

[12] For more information about Optuna: https://optuna.org/.

elastic net considers residuals for all data observations, the loss function of SVR only takes into account a subset of data observations within and on its support vectors. Statistically, LASSO and ridge regression are special cases of the elastic net when $\lambda = 1$ and $\lambda = 0$. To construct a pure out-of-sample signal, at each point in time t, we use a rolling window of the most recent three months' text data to fit the model above and obtain the coefficients of α_t and β_t^k. Similar to SVR, we first fit the text data using the elastic net method to obtain estimates of α_t and β_t^k. We then use the fitted values from the model to construct the predicted delta-hedged option returns based on textual predictors:

$$\hat{r}_{i,t+1} = \hat{\alpha}_t + \sum_{k=1}^{K} \hat{\beta}_t^k x_{i,t}^k, \qquad i = 1, \ldots, N_t. \tag{10}$$

Another difference between elastic net and SVR is that elastic net can shrink some coefficients to zero (i.e., $\hat{\beta}_t^k = 0$), thus the model may have a sparse structure compared to SVR. Therefore, it is easier to determine the feature importance under the elastic net. Although SVR and elastic net can select the most relevant textual information from news media, they do not allow nonlinearity and interactions among predictors, which are likely useful for predicting option returns using textual information because words are dependent on each other. To incorporate nonlinearity and interactions among words, we consider more advanced machine-learning approaches such as random forest and neural networks. The recent study by Gu, Kelly, and Xiu (2020) shows that these methods are helpful in forecasting stock returns using quantitative signals.

Random forest regression is a powerful ensemble method that combines multiple decision trees to improve prediction accuracy and reduce the overfitting problem. The random forest regression is conducted in three steps: from the full sample data S, we first draw a subsample with replacement $\{S^b\}_{b=1}^B$ that has n observations and m randomly sub-selected features. Second, we train a decision tree and obtain a predictor $\hat{r}^b$ on each S^b. Finally, we take the average among all subsamples with sub-selected features:

$$\hat{r}^{RF}(x) = B^{-1} \sum_{b=1}^{B} \bar{r}(T_b^*(x)), \tag{11}$$

where $T_b^*(x)$ denotes a random-forest tree with bootstrapped data and sub-selected features, and x is a certain predictor.

For the neural network, we use a version of feed forward network, also known as multilayer perceptron (MLP) regression. The units in the MLP regression are arranged into a set of layers, and each layer contains some number of identical units with a pre-specified activation

function such as the softmax function (Softmax), rectified linear activation (ReLU), the logistic activation (Sigmoid), and the hyperbolic tangent activation (Tanh). Every unit in each layer is connected to every unit in the next layer. The first layer is the input layer, while the last one is the output layer, which is a single unit in our case. All layers in between are defined as hidden layers. To fix the idea, consider a simple case with two consecutive layers. The network's computations can be written as:

$$\begin{aligned} h_i^{(1)} &= \phi^{(1)}\Big(\sum_j w_{ij}^{(1)} x_j + b_i^{(1)}\Big), \\ h_i^{(2)} &= \phi^{(2)}\Big(\sum_j w_{ij}^{(2)} h_j^{(1)} + b_i^{(2)}\Big), \\ r_i &= \phi^{(3)}\Big(\sum_j w_{ij}^{(3)} h_j^{(2)} + b_i^{(3)}\Big). \end{aligned} \tag{12}$$

The nonlinearity and interaction among words can be captured by the nonlinear activation functions and the full connections among the hidden layers. Under the Universal Approximation Theorem (Cybenko (1989) and Hornik, Stinchcombe, and White (1989)), a neural network with one hidden layer can approximate any continuous function for inputs within a specific range. For robustness concerns, we consider different numbers of hidden units and neuron sizes.[13]

[Insert Table 5]

To save space, Table 5 presents the single portfolio sorting results of each textual predictor trained by alternative machine-learning approaches. The results of regressions are similar and available upon request. We show that the textual information from news media extracted by different machine-learning approaches can significantly and robustly predict delta-hedged equity option returns. It is important to note that our analysis confirms the robustness of our results across different machine-learning methods, but it is not designed to compare the efficacy of these various models. The alternative ML textual predictors are highly correlated with the SVR textual predictors, suggesting that different ML approaches capture similar useful information from news media. Table A6 shows the correlations of textual predictors across different ML approaches. For call options, the elastic net textual predictors have a correlation coefficient of 0.70 with the SVR textual predictors, while the neural network and

[13]Following suggestions in Gentzkow, Kelly, and Taddy (2019) that it is more beneficial to apply complex models after some dimension reduction, we shrink the 10,000-dimensional $tf-idf$ matrix to a 100-dimensional matrix using the principal component analysis (PCA) and input this resulting matrix into the SVR model.

random forest textual predictors have correlation coefficients of 0.48 and 0.66 with the SVR textual predictors, respectively.

3.2.2 Alternative Constructions of Word Features

In our main analysis, we train the SVR model using unigram word counts (adjusted by document frequency), because of its simplicity and effectiveness. However, it has two limitations. First, it ignores word dependency in different contexts. The semantic meaning of a unigram feature may vary depending on the adjacent word. Second, it reduces interpretability. Some unigram features only make sense when paired with other words, such as collocations and noun phrases. A possible solution is to use features with more than one word, such as bigrams or trigrams. For instance, a bigram feature is the combination of two consecutive words in a sentence.

By constructing features in n-grams, we are able to mitigate the semantic differences caused by word dependency, thereby enhancing model interpretability compared to the unigram feature approach. To check whether our empirical findings are robust to other choices of word features, we retrain our SVR model to forecast equity option returns using various n-gram features. We consider bigrams, trigrams, and the combination of unigrams and bigrams. We treat each n-gram as a new feature and use the $tf - idf$ process to adjust their counts. The n-gram features are then used to train the SVR model specified in Equation (4) and construct the corresponding textual predictor based on Equation (5). Given that the total number of features increases exponentially when switching from unigrams to n-grams ($n > 1$), we include more features in the input of the SVR model. Specifically, 40,000 (80,000) features are input to the SVR model for the bigram (trigram) case. For the combination of unigrams and bigram, we input 20,000 features into the SVR model. The empirical results are provided in Table 6. To save space, Table 6 presents the single portfolio sorting results of each textual predictor trained by alternative word constructions.

[Insert Table 6]

Table 6 reports the predictive power of textual predictors from news media with different n-gram features for the cross-section of delta-hedged equity option returns. A larger n for the n-gram feature (e.g., bigram or trigram) may enhance the interpretability of the textual predictors, but it may also introduce more noises and computational burden. The high computational costs limit the number of word features that we can include for bigrams or trigrams, leading to information loss. For example, the top 40,000 bigram tokens cover 24.32% of the total bigram term frequencies, and the top 80,000 trigram tokens cover only

9.29% of the total trigram term frequencies, as shown in Table A2. Therefore, the return spreads in Table 6 may be smaller than those in our main results, however, they are still significant and robust. Notably, combining unigrams and bigrams, and incorporating more word features, does enhance our results to some degree. However, our analysis shows that using unigrams alone already yields satisfactory results. The incremental benefit of adding bigrams and more features is not substantial. Therefore, we opt to focus on unigrams in our main analysis.

3.2.3 Word Embedding and Large Language Model

In our baseline analysis, we implement the term frequency-inverse document frequency technique ($tf - idf$) to construct the feature matrix. However, as a type of Bag-of-Words (BOW) model, it may be considered somewhat outdated and limited in capturing the semantic relationships between words. In this subsection, we employ more advanced text representation techniques, such as word embeddings and large language models (LLM), to represent our text data and test whether our results are robust to different techniques chosen to represent the text data.

Word embedding is a type of word representation that encodes the meaning of words such that words closer in the vector space are likely to have similar meanings. To implement this method, we employ the Word2Vec approach proposed by Mikolov, Chen, Corrado, and Dean (2013). Word2Vec represents a word as a high-dimensional vector of numbers, capturing relationships between words. A famous illustrative example that demonstrates the efficiency of Word2Vec is the operation of vector("King") - vector("Man") + vector("Woman"), which results in a vector aligning most closely with the vector representation of the word "Queen". To implement Word2Vec, we use pre-trained word vectors to tokenize our text data. Following Chen et al. (2023), we download and implement the model "wiki-news-300d-1M", which contains 1 million 300-dimensional word vectors trained on Wikipedia and large corpus of news articles.[14] Specifically, for a token in an article, it is transformed into a 300-dimensional vector, and we take the average of all tokens' vectors in an article to obtain the article-level embedding. For each firm in each month, we then take the average of the article-level embeddings to derive the embeddings at the firm-month level.

In addition to word embeddings, we apply state-of-the-art large language models to tokenize our text data. Specifically, we use the widely adopted BERT and FinBERT models to generate article representations. BERT is a large language model based on the novel transformer architecture, first introduced by Google in October 2018. It is pre-trained on a

[14]The model can be downloaded from the fastText.

large corpus of text from sources like Toronto BookCorpus and English Wikipedia, enabling it to learn a vast range of linguistic patterns and contextual nuances, resulting in significant improvements over previous models. FinBERT, a domain-specific adaptation of BERT, is tailored for financial sentiment analysis. It is pre-trained on a financial corpus and fine-tuned for financial sentiment analysis. Compared with the Word2Vec approach, which is based on a 2-layer shallow neural network architecture and can only capture contextual information within a small fixed-size window, large language models based on the transformer architecture, such as BERT and FinBERT, utilize an encoder design of deep neural networks. These models, pre-trained on extensive corpora, offer contextualized embeddings that encapsulate nuanced semantic information and contextual dependencies within the text. Since BERT and FinBERT can process input sequences up to 512 tokens, we focus on the initial 512 tokens for articles exceeding this limit. BERT and FinBERT then translate these tokens into 768-dimensional vector representations. We take the average of the token-level representations to obtain the article-level representation. For each firm in each month, we then average the article-level representations to derive the firm-month-level representations.

After obtaining vector representations for our text data, we input these representations into the SVR model to derive the corresponding textual predictors, ensuring consistency with our baseline analysis. Table 7 shows that the predictability of textual predictors based on word embeddings and large language models (LLM) is also significant and robust. Among these, the BERT model performs particularly well. The return spread obtained by textual predictors based on BERT is 0.61% for call options and 0.42% for put options, showing an improvement compared to the return spreads in our baseline analysis, though the improvement is not striking. The performance of word embeddings and FinBERT, while slightly weaker, is still very significant and comparable to our baseline results in magnitude. Furthermore, it is worth noting that our baseline results, which are based on the Bag-of-Words method and the $tf - idf$ transformation, already perform very well. This suggests that the deep language understanding captured by more complex text representations may not be crucial for our downstream task: forecasting option returns. Instead, individual key terms are highly indicative of future option returns. This motivates us to construct dictionaries for equity option returns, and thanks to the transparency and interpretability provided by the Bag-of-Words method and the linear SVR model, we construct these dictionaries in Section 4.

[Insert Table 7]

3.2.4 Time Series of Return Spreads and Subperiod Analysis

Avramov, Cheng, and Metzker (2023) documents that trading strategies in the stock market utilizing machine learning yield higher profits in high limits-to-arbitrage market states, such as high investor sentiment, high market volatility, and low market liquidity. In this section, we conduct a subperiod analysis to examine the robustness of the return spreads generated by the SVR textual predictors in different market conditions. In particular, we consider four criteria: sentiment, volatility, liquidity, and recession phases.

We use the Baker and Wurgler (2006) sentiment index to distinguish between periods of high and low market sentiment. A period is classified as having high sentiment when its sentiment index is higher than the median sentiment index for the entire sample period. We measure the market-wide volatility using the CBOE VIX index. As for the market-level liquidity, we use the equal-weighted Amihud (2002) stock illiquidity measure of individual stocks as a proxy for the overall market liquidity. Lastly, recession periods in our sample are identified based on the recession timelines provided by the National Bureau of Economic Research (NBER).

This subperiod analysis allows us to evaluate the performance of SVR textual predictors under different market conditions. Table 8 shows that the return spreads generated by SVR textual predictors are robust in different market conditions. Furthermore, aligning with findings of Avramov et al. (2023), we observe that the return spreads are significantly higher during periods with high investor sentiment, high market volatility, and low market liquidity. For example, the difference in return spreads in periods with high investor sentiment is 0.15% (0.13%) higher than in periods with low investor sentiment for call (put) options, showing a 26.79% (36.11%) difference compared with the return spreads in the entire sample.In addition, the return spreads are also significantly higher during recession periods.

[Insert Table 8]

3.2.5 Predictability for Returns of Alternative Portfolios

In addition to testing the predictability of news data information on returns of delta-hedged option portfolios, we also investigate whether textual predictors obtained from the news media corpus using the SVR model have predictive power for returns of alternative option portfolios. In this subsection, we consider three different types of straddle portfolios: vanilla straddle as described in Goyal and Saretto (2009), zero-beta straddle as described in Coval and Shumway (2001), and delta-neutral straddle as described in Gao, Xing, and

Zhang (2018). Furthermore, we consider a skewness portfolio, as defined in Bali and Murray (2013), which primarily has exposure to skewness.

To construct straddle portfolios, at the end of each month, we select pairs of call and put options on the same underlying stocks that are closest to being at-the-money and have the same strike price and maturity, using the option sample after applying the standard filters mentioned in Section 2.1. The vanilla straddle portfolio consists of purchasing one contract of a call option and one contract of a put option, respectively. Its return for firm i from t to $t+1$ is defined as follows:

$$r_{i,t}^{Vanilla\ \ Straddle} = \frac{C_{i,t+1} + P_{i,t+1}}{C_{i,t} + P_{i,t}} - 1, \tag{13}$$

where $C_{i,t}$ ($P_{i,t}$) is the price of call (put) option for firm i at time t.

Following Coval and Shumway (2001), we consider the returns of zero-beta straddle portfolios, which are beta-neutral and have zero exposure to market risk. Specifically, the zero-beta straddle returns are computed as the weighted average of raw returns of call and put options as follows:

$$r_{i,t}^{Zero-beta\ \ Straddle} = w_{i,t} r_{i,t}^{Call} + (1 - w_{i,t}) r_{i,t}^{Put}, \tag{14}$$

where $r_{i,t}^{Call}$ ($r_{i,t}^{Put}$) is the raw return of call (put) options for firm i at time t. The weight $w_{i,t}$ is given by:

$$w_{i,t}^{Zero-beta} = \frac{C_{i,t}(1 - \Delta_{i,t}^{Call})}{P_{i,t}\Delta_{i,t}^{Call} + C_{i,t}(1 - \Delta_{i,t}^{Call})} \tag{15}$$

where $\Delta_{i,t}^{Call}$ is the delta of the call option for firm i at time t.

Moreover, following Gao et al. (2018), we also consider the returns of delta-neutral straddle portfolios to neutralize the directional exposure to the underlying stocks. The delta-neutral straddle returns are also computed as the weighted average of the raw returns of the call and put options, with the weight given by:

$$w_{i,t}^{Delta-neutral} = -\frac{\Delta_{i,t}^{Put}}{\Delta_{i,t}^{Call} - \Delta_{i,t}^{Put}}. \tag{16}$$

Delta-hedged option portfolios and straddle portfolios primarily have exposure to the implied volatility, which is the second moment of the risk-neutral distribution. In addition to them, we consider the skewness asset constructed in Bali and Murray (2013) to isolate the effect of volatility while maintaining exposure to skewness. Specifically, the skewness

portfolio is constructed by purchasing one contract of an out-of-the-money call option, w_1 contracts of out-of-the-money put options, and w_2 contracts of the underlying stocks.[15] The values of w_1 and w_2 are determined by the following equation:

$$\begin{aligned} w_1 &= -\frac{V_C}{V_P} \\ w_2 &= \Delta_C + w_1 \Delta_P, \end{aligned} \tag{17}$$

where V_C (V_P) is the vega of the call (put) option, and Δ_C (Δ_P) is the delta of the call (put) option. This skewness portfolio effectively hedges away the effects of implied volatility of the underlying stocks (vega-neutral) and directional changes (delta-neutral) of the underlying stocks, while maintaining exposure to the skewness.

[Insert Table 9]

Table 9 shows that textual predictors for returns of alternative portfolios exhibit strong predictive power. The return spread for vanilla straddle portfolios sorted by textual predictors is 2.17% with t-statistics of 3.12. The return spreads for delta-neutral and zero-beta straddle portfolios are 4.48% with t-statistics of 5.38 and 2.78% with t-statistics of 4.17, respectively. The return spread for the skewness portfolios is also significant, with a magnitude of 0.36% and t-statistics of 3.40. These return spreads are significant both statistically and economically, indicating that the predictability of signals extracted from news media data is not limited to specific option portfolios.

4 Interpretations of Textual Predictors and Information Contents

4.1 Nature of the Textual Information

4.1.1 Important Words in Constructing Textual Predictors

We have presented extensive evidence indicating that qualitative information sourced from the news media contributes valuable insights to forecasting delta-hedged option returns. However, the economic mechanism behind such return predictability remains uncertain, particularly considering that its extraction relies primarily on intricate machine learning models.

[15] Out-of-the-money call (put) option is defined as the call (put) option with a delta closest to 0.1 (-0.1).

In this section, we propose a novel method to offer understandings of the interpretation of the SVR textual predictors.

To explore this question, we first create a dictionary that captures key features from option returns data. This dictionary serves as an intuitive means to visually represent textual information pertinent to delta-hedged option returns. In our methodology, we sort word features in the SVR model into positive and negative groups based on the sign of their coefficients during each training iteration. From each group, we select the top 2,000 words with the largest absolute value of the magnitude of the coefficients as the important words. Subsequently, we compile and count these words' occurrences, noting their positive or negative impact on option returns. We then define two scores to organize the resulting datasets:

$$\begin{aligned} \text{Positive Score} &= \frac{P}{P+N} \\ \text{Negative Score} &= \frac{N}{P+N}, \end{aligned} \tag{18}$$

where, P represents the count of occurrences that a specific word is positively associated with delta-hedged option returns, while N denotes the number of occurrences that the same word is negatively associated with delta-hedged option returns. The sum $P+N$ indicates the total frequency of this word being selected as important words. To filter out rare words, we set a threshold based on the frequency of word occurrence in our analysis. Specifically, since our analysis involves 51 rolling training iterations, a word must appear at least 25 times to be included in our dictionary. Consequently, for both call and put options, we form two distinct dictionaries related to the delta-hedged option returns: a positive and a negative dictionary. For brevity, Table A7 displays the top 100 terms with the highest frequency of positive or negative occurrences from each dictionary, with the full version available on the authors' website for replication and extended use. Furthermore, we show the top 100 bigrams in Tables A8. Figure 1 presents word clouds for each dictionary, where the font size of each word corresponds to the frequency of its association with delta-hedged option returns.

[Insert Figure 1]

We find that the overlap between important words that are positive (negative) in the call option dictionary and the put option dictionary is 58.6% (56.5%). In contrast, only 1.9% of the positive words in the call option dictionary overlap with the negative words in the put option dictionary, and 2.3% vice versa. This suggests that the majority of word

features convey information affecting both call and put options similarly, indicating that the effectiveness of our textual predictors is not attributable to the underlying stock returns' drift term. A significant overlap of words with opposing signs would have been expected if the underlying stock returns drive the return predictability.

From the dictionary we construct for equity option returns, we can gain some interesting economic insights. We focus on words that are related to both call and put option returns in the same direction. For example, many words that are positively related to option returns are associated with macroeconomics or commodities, which have market-wide impacts. In the positive option return dictionaries, terms such as *inflation*, *policy*, *economic growth*, *interest rate*, *supply chain*, and *central bank*, *presidential campaign* are clearly related to macroeconomic conditions. In addition, terms such as *energy*, *barrel*, *metal*, *gold*, and *mine*, *energy price*, *gold price* are related to commodities. Fluctuations in macroeconomic conditions or commodity prices can significantly impact various industries and firms, resulting in increased market volatility. Companies discussed in articles that frequently mention macroeconomic conditions tend to be more sensitive to these changes, leading to increased future volatility and higher equity option returns.

In contrast, words that exhibit a negative correlation with option returns are closely associated with financial stress or specific corporate issues. For example, terms such as *bankruptcy*, *debt*, *divestiture*, *restructure*, *lawsuit*, *investigation*, *violation*, *bankruptcy protection*, *file bankruptcy*, *merger agreement*, and *agree acquire* are negatively related to equity option returns. Our observation aligns with the findings of Vasquez and Xiao (2024), which find that default risk can significantly and negatively forecast future delta-hedged option returns. They argue that investors are willing to pay a premium to hedge against the variance increases in firms with high default risk. Furthermore, we can apply the similar reasoning to explain why many words are associated with crucial corporate issues: investors use equity options to hedge against the increased uncertainty arising from these events and are willing to pay premiums for equity options.

In addition, words that are negatively related to equity option returns demonstrate significant industry concentration, particularly within the pharmaceutical sector. For example, in the negative option return dictionary, terms such as *trial*, *treatment*, *patient*, *clinical*, *drug*, *develop drug*, *clinical trial*, *drug development*, *approve drug*, *cancer treatment*, and *experimental drug* stand out. Once a drug receives approval, the associated pharmaceutical company often experiences substantial price appreciation. Consequently, investors are willing to pay a premium for equity options written on these companies. This observation is consistent with the findings of Andreou, Bali, Kagkadis, and Lambertides (2023), which suggests that option investors tend to overreact to the high growth potential.

4.1.2 Topic Analysis of Option Return Dictionaries

Although important word features offer some insights, they can be too detailed and not easy to understand at first. This section aims to sort these word features into more easily interpretable topics. This approach offers us a clearer picture of the key themes that drive the predictability of textual information. Bybee et al. (2021) utilize the Latent Dirichlet Allocation (LDA) model to categorize words into 180 topics and assign a vector of weights to each word for each topic based on its relevance to each topic.[16] We apply these weights to our option return dictionary. Specifically, for each word in our option return dictionary, we obtain its vector of weights for the 180 topics and then aggregate these weights for each dictionary by summing up the weights across words. We rank the topics according to their total weights from highest to lowest and present the top 10 of them in Figure 2.[17]

Figure 2 presents the topic classification for our option return dictionary, offering insightful observations. It appears that words positively associated with option returns relate to broader, often industry- or market-level topics. For instance, topics like *Product prices*, *Rental properties*, *Machinery*, *Small business*, *Oil market*, *Steel*, *Economic growth*, and *Mining* show positive correlations with both call and put option returns. Conversely, financial distress or event-specific topics tend to negatively impact option returns. In particular, among the top 10 topics, *Bankruptcy*, *IPOs*, *Earnings forecasts*, *Share payouts*, and *Earnings losses* are negatively associated with future equity option returns.

[Insert Figure 2]

4.2 Information Contents of the Predictability

4.2.1 Word Overlap Analysis

In this section, we employ both qualitative and quantitative methods to explore the sources that significantly contribute to the predictive power of our SVR textual predictors. Table 4 shows that the inclusion of various ex-ante determinants of expected option returns leads to a substantial decrease in the coefficients of our SVR textual predictors (about one-third for both call and put options), indicating some overlap in the information content of the textual predictors and these option return determinants. We hypothesize that news

[16] The weight of term v for topic k is determined by the number of times term v is assigned to topic k (the ϕ parameter in Bybee et al. (2021)). We use the absolute word weights instead of the scaled word weights.

[17] We thank authors of Bybee et al. (2021) for providing their data on their website: http://structureofnews.com/#

articles, characterized by their use of words, contain information related to several important option return determinants, and a significant portion of the predictive power of news data for delta-hedged option returns stems from such information. Figure 3 presents a causal diagram that illustrates this concept.

[Insert Figure 3]

To study this question, we analyze the overlap between the words that significantly forecast delta-hedged option returns in SVR and words that are related to various option return determinants including changes in future implied volatility (ΔIV), idiosyncratic volatility (IVOL), volatility deviation (HV–IV), Amihud stock illiquidity measure (ILLIQ), uncertainty of implied volatility (VOIV), model-free implied skewness (MFIS), and model-free implied kurtosis (MFIK). The assumption is that each option return determinant has the corresponding information set in the news data corpus. Thus, we should be able to identify the information overlap between delta-hedged option returns and various determinants based on their corresponding dictionaries constructed by projecting each quantitative variable onto the space of the news data corpus. Accordingly, we conduct the analysis as follows: first, we use the SVR algorithm to identify the important words that have positive or negative effects on each option return determinant, which is used as the target variable. We then obtain two lists of important words, namely the positive and negative lists, for a certain option return determinant, each containing 1,000 words. We consider these word lists as the positive and negative dictionaries for such option return determinant. We take the negative value of a determinant as the target variable if it negatively predicts delta-hedged option returns, such that all these features have positive relations with delta-hedged option returns, consistent with the direction of our option return dictionary. We use these derived dictionaries to proxy for the information set related to each option return determinant. Like our option return dictionary, we list in Table A9 the top 100 words of positive and negative dictionaries for each option return determinant. The whole dictionary for each option determinant is available on the authors' website for public usage.

Second, we compare each dictionary with the corresponding positive or negative option return dictionary that we obtain in Section 4.1.1. For example, we can assign each word in the positive dictionary for idiosyncratic volatility to one of these two groups: 1) words that are positively related to both idiosyncratic volatility and delta-hedged option returns; 2) words that are positively related to idiosyncratic volatility but negatively related to delta-hedged option returns. Since we reverse the sign of idiosyncratic volatility and make it positively correlated with delta-hedged option returns, we consider the first group as the correct overlap and the second group as the incorrect overlap. We repeat this process for the

negative dictionary for idiosyncratic volatility. Finally, we define an overlap score for each option return determinant that quantifies its degree of correct word overlaps. The overlap score is given by:

$$\text{Overlap Score} = \frac{C - W}{C + N + W} \tag{19}$$

where C is the number of words (among the top 2,000 words) that overlap correctly, N is the number of non-overlapping words, and W is the number of words that overlap incorrectly. Figure 4 illustrates the calculation of the overlap score between the dictionary for changes in implied volatility and the dictionary for call option returns. The overlap score measures how well the information of a given option return determinant can help us classify a word into the correct option return dictionary. It reflects the similarity between the information sets of a certain option return determinant and delta-hedged option returns. Essentially, a higher overlap score means more shared information sets between the option return determinant and delta-hedged option return, while a lower score indicates less shared information sets. For instance, our analysis reveals that the overlap score between the call and put option dictionaries is notably high at 55.45%.

[Insert Figure 4]

[Insert Table 10]

We calculate the overlap score between option return dictionaries and six option return determinant variables, and present the results in Table 10. Among various option return determinants, the highest average overlap score is observed for the changes of implied volatility, which has an average overlap score of 21.8% (21%) for the call (put) option dictionary.[18] This suggests that the information set related to changes of implied volatility is a key source of textual information for option return predictability. Other option return determinants that have relatively high overlap scores include idiosyncratic volatility (20.9% for call and 19.7% for put), model-free implied skewness (17.71% for call and 13.03% for put), and model-free implied kurtosis (10.6% for call and 8.5% for put), which capture higher moment information (e.g., jump risk) and market frictions. These results highlight the importance of jump risk and market friction-related information in shaping textual predictors for option returns. Furthermore, the Amihud illiquidity measure and the uncertainty of implied volatility also

[18] We confirm this finding in a time series setting. Specifically, in each training iteration, we compute the overlap score for every option return determinant and draw the pattern of these overlap scores. Figure A1 shows the time series pattern of overlap scores, with the overlap score of the changes of implied volatility consistently ranking the highest most of the time.

exhibit meaningful overlap scores, indicating their contribution to the predictability of the textual information.[19]

Table 10 shows the advantage of the SVR textual predictors in capturing information from different aspects of the news data corpus, many of which are difficult to quantify using a lexicon-based approach. Besides information related to implied volatility changes, market frictions, and jump risks, we find that information related to stock illiquidity and uncertainty about implied volatility also contribute to the predictive power of the textual predictors. However, it is notable that HV–IV, despite being a key predictor of delta-hedged option returns, does not show a significant overlap with the information sets of option returns present in the news corpus.

4.2.2 Decomposition Analysis

In this section, we take a step further and test which option return determinants contribute the most to the textual option return predictor. Given the news data as the whole information set, we project each option return determinant onto the news corpus space and obtain the projected value for these determinants. In particular, we use each of the option return determinants as the target variable and use the SVR to get the corresponding fitted value based on all the word features we used to train option return models. This approach ensures that the information set is confined to the news data corpus and is independent of other data sources. After obtaining these fitted values, we implement the Fama-Macbeth regression method to decompose the contributions of these determinants to the SVR option return predictor:

$$\mathrm{TP}_{i,t} = \hat{\alpha}_t + \sum_{k=1}^{K} \hat{\beta}_t^k x_{i,t}^k, \quad i = 1, \ldots, N_t \tag{20}$$

where $TP_{i,t}$ is the SVR textual predictor for firm i in month t, and $x_{i,t}^k$ is the projected value of a certain option return determinant on the news data corpus. The independent variables are standardized to have zero mean and one standard deviation, so that the magnitudes of their coefficients are comparable.

[Insert Table 11]

Table 11 shows that changes of implied volatility has the highest coefficient and adjusted

[19] We confirm the validity of our word overlap method by applying it to the returns of the skewness portfolio described in Section 3.2.5. We find that, among option return determinants, model-free implied skewness has the highest overlap score with the dictionary of the skewness portfolio return, and its overlap score is significantly higher than those of other determinants.

R-squared, implying that it is the most important source of textual information contributing to the textual predictor. Idiosyncratic volatility, illiquidity, model-free implied skewness or kurtosis, and uncertainty of implied volatility are also substantial contributors, while volatility deviation and HV-IV have wrong signs (for call options) or no significant effect (for put options). In addition to this, we also compare the effects of all the fitted option return determinants using a horse-racing regression. In the last column, we confirm that changes of implied volatility is the most influential factor that contributes to the textual predictor. Furthermore, the coefficients of those aforementioned option return determinants remain significant.[20] This result highlights the advantage of applying machine learning approaches compared to traditional lexicon-based methods, as a machine learning model can incorporate a combined effect across multiple predictive resources for option return predictability, especially for some variables that are difficult to quantify through a predefined dictionary. While information in news contents related to changes of implied volatility is the major source of the predictability of our textual predictors, it is noteworthy that the fitted values of changes of implied volatility do not exhibit comparable predictability for delta-hedged option returns. In an untabulated table, we verify that the fitted values of the above-mention option return determinants do not match the level of predictability provided by our textual predictors for option returns. Our results thus highlight the importance of using delta-hedged option returns as the target variable to train the model.

4.2.3 Generalizability to Earnings Conference Call Transcripts

Although the textual predictors obtained from the SVR model applied to the news media corpus perform well in predicting delta-hedged option returns, it remains a question whether the trained SVR model works well when applied to other textual corpora. In this section, motivated by Jegadeesh and Wu (2013), we investigate the generalizability of our SVR model trained on the news media corpus to another important text data, earnings conference call transcripts. Earnings conference calls have become an increasingly important medium through which firms' management teams disseminate information to the market. Recent literature has demonstrated the importance of the information conveyed in earnings conference calls. For example, Hassan, Hollander, van Lent, and Tahoun (2019) and Sautner, Van Lent, Vilkov, and Zhang (2023) construct measures for firm-level political risk and climate change exposure from earnings conference call transcripts, respectively. Applying the

[20]The sign of the fitted value of the model-free implied kurtosis flipped because it captures similar information with the model-free implied skewness. Similarly, since stock illiquidity captures similar information with idiosyncratic volatility and the model-free implied skewness, its coefficients flip sign for call options and become insignificant for put options. The sign of HV-IV is more negative in the horse-racing specification, reassuring its little contribution to the formation of our textual predictors.

SVR model trained on the news media corpus to earnings conference call transcripts allows us to perform an out-of-sample test to confirm the validity and reliability of our model.

Our sample of earnings conference call transcripts obtained from Capital IQ Transcripts covers more than 8,000 public companies and includes both the executives' presentations and Q&A sessions. After merging with option returns and control variables, we have 55,881 observations spanning from November 2005 to December 2021. We then clean and transform the dataset into a $tf-idf$ matrix as in our baseline analysis.[21] Next, we apply the trained SVR model from the news media corpus to this $tf-idf$ matrix of the earnings conference call transcripts to get the new textual predictors. Since we train our SVR model on the news media corpus in a rolling style, we apply each corresponding model to the time-matched sample of the earnings conference call transcripts, ensuring no look-ahead bias. Specifically, the textual predictors obtained from earnings conference call transcripts are calculated as follows:

$$TP_{i,t}^{Conf} = \hat{\alpha}_t + \hat{\beta}_t' x_{i,t}^{Conf}, \tag{21}$$

where $x_{i,t}^{Conf}$ is the transformed $tf-idf$ matrix of the earnings conference call transcripts, and $\hat{\alpha}_t$ and $\hat{\beta}_t'$ are parameters of the SVR model trained on the news media corpus at time t.

[Insert Table 12]

Due to the strong seasonality in the scheduling of earnings conference calls, which results in too few observations for certain months, we use panel regression with time fixed effects and cluster standard errors by time. The results presented in Table 12 show the generalizability of the SVR model trained on the news media corpus. $TP_{i,t}^{Conf}$ shows significant and robust predictability for delta-hedged option returns, and this predictability cannot be subsumed by control variables included in our baseline analysis. In untabulated results, we find that both sentiment and uncertainty measures derived from the LM dictionary using earnings conference call transcripts lack predictive power for delta-hedged option returns on their own.

5 Conclusion

In this paper, we study whether and how textual information from the news media can be used to enhance option return predictability. First, we find that the news media contains

[21] It is worth noting that the $tf-idf$ vectorizer is trained on the news media corpus.

substantial information for future delta-hedged option returns. The results are robust after controlling for quantitative option return predictors documented in the literature. Furthermore, our results are robust to the choices of machine-learning algorithms and different ways to construct word features. Our results demonstrate that the news media contain qualitative information useful for option return prediction.

It is interesting but challenging to pin down the underlying mechanisms for the option return predictability by textual information from news media. In this paper, we propose a novel method to answer this question. Employing both qualitative and quantitative methods, we find that the predictive power of the textual predictors arises from a composite effect, with changes in future implied volatility being the most influential, followed by idiosyncratic volatility and implied skewness. Future research could also extract textual indicators from other types of alternative data to forecast equity option returns, such as earnings conference calls, analyst reports, and Federal Reserve press conference transcripts. Another direction could be exploring more advanced machine learning approaches (such as recurrent neural network and convolutional neural network) and incorporating word dependency across words in a document in order to extract contextual information from text data.

References

Amihud, Yakov, 2002, Illiquidity and stock returns: Cross-section and time-series effects, *Journal of Financial Markets* 5, 31–56.

Andreou, Panayiotis C., Turan G. Bali, Anastasios Kagkadis, and Neophytos Lambertides, 2023, Firm growth potential and option returns, *Working Paper* .

Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, The cross-section of volatility and expected returns, *Journal of Finance* 61, 259–299.

Avramov, Doron, Si Cheng, and Lior Metzker, 2023, Machine learning vs. economic restrictions: Evidence from stock return predictability, *Management Science* 69, 2587–2619.

Baker, Malcolm, and Jeffrey Wurgler, 2006, Investor sentiment and the cross-section of stock returns, *Journal of Finance* 61, 1645–1680.

Bakshi, Gurdip, and Nikunj Kapadia, 2003, Delta-hedged gains and the negative market volatility risk premium, *Review of Financial Studies* 16, 527–566.

Bali, Turan G., Heiner Beckmeyer, Mathis Mörke, and Florian Weigert, 2023, Option return predictability with machine learning and big data, *Review of Financial Studies* 36, 3548–3602.

Bali, Turan G., and Scott Murray, 2013, Does risk-neutral skewness predict the cross-section of equity option portfolio returns?, *Journal of Financial and Quantitative Analysis* 48, 1145–1171.

Boulatov, Alex, Assaf Eisdorfer, Amit Goyal, and Alexei Zhdanov, 2022, Limited attention and option prices, *Working Paper* .

Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu, 2021, The structure of economic news, *Journal of Finance, forthcoming* No. 26648.

Cao, Jie, and Bing Han, 2013, Cross section of option returns and idiosyncratic stock volatility, *Journal of Financial Economics* 108, 231–249.

Cao, Jie Jay, Aurelio Vasquez, Xiao Xiao, and Xintong Eunice Zhan, 2023, Why does volatility uncertainty predict equity option returns?, *Quarterly Journal of Finance* 13.

Chen, Yifei, Bryan T. Kelly, and Dacheng Xiu, 2023, Expected returns and large language models, *Working Paper* .

Chinco, Alex, Adam D. Clark-Joseph, and Mao Ye, 2019, Sparse signals in the cross-section of returns, *Journal of Finance* 74, 449–492.

Choy, Siu Kai, and Jason Wei, 2023, Investor attention and option returns, *Management Science* .

Cortes, Corinna, and Vladimir Vapnik, 1995, Support-vector networks, *Machine Learning* 20, 273–297.

Coval, Joshua D., and Tyler Shumway, 2001, Expected option returns, *Journal of Finance* 56, 983–1009.

Cybenko, G., 1989, Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals, and Systems* 2, 303–314.

Dong, Xi, Yan Li, David E Rapach, and Guofu Zhou, 2021, Anomalies and the expected market return, *Journal of Finance* 77, 639–681.

Engle, Robert F., Stefano Giglio, Bryan Kelly, Heebum Lee, and Johannes Stroebel, 2020, Hedging climate change news, *Review of Financial Studies* 33, 1184–1216.

Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Fama, Eugene F., and James D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy* 81, 607–636.

Frankel, Richard, Jared Jennings, and Joshua Lee, 2022, Disclosure sentiment: Machine learning vs. Dictionary methods, *Management Science* 68, 5514–5532.

Gao, Chao, Yuhang Xing, and Xiaoyan Zhang, 2018, Anticipating uncertainty: Straddles around earnings announcements, *Journal of Financial and Quantitative Analysis* 53, 2587–2617.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, 2019, Text as data, *Journal of Economic Literature* 57, 535–574.

Goyal, Amit, and Alessio Saretto, 2009, Cross-section of option returns and volatility, *Journal of Financial Economics* 94, 310–326.

Gu, Shihao., Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *Review of Financial Studies* 33, 2223–2273.

Hassan, Tarek A., Stephan Hollander, Laurence van Lent, and Ahmed Tahoun, 2019, Firm-level political risk: Measurement and effects, *Quarterly Journal of Economics* 134, 2135–2202.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White, 1989, Multilayer feedforward networks are universal approximators, *Neural Networks* 2, 359–366.

Huang, Darien, Christian Schlag, Ivan Shaliastovich, and Julian Thimme, 2019, Volatility-of-volatility risk, *Journal of Financial and Quantitative Analysis* 54, 2423–2452.

Jegadeesh, Narasimhan, and Di Wu, 2013, Word power: A new approach for content analysis, *Journal of Financial Economics* 110, 712–729.

Jeon, Yoontae, Raymond Kan, and Gang Li, 2024, Stock return autocorrelations and the cross section of option returns, *Management Science, forthcoming* .

Jeon, Yoontae, Thomas H. McCurdy, and Xiaofei Zhao, 2022, News as sources of jumps in stock returns: Evidence from 21 million news articles for 9000 companies, *Journal of Financial Economics* 145, 1–17.

Ke, Zheng Tracy, Bryan T. Kelly, and Dacheng Xiu, 2019, Predicting returns with text data, *National Bureau of Economic Research Working Paper Series* No. 26186.

Kelly, Bryan, Asaf Manela, and Alan Moreira, 2021, Text selection, *Journal of Business & Economic Statistics* 39, 859–879.

Loughran, Tim., and Bill McDonald, 2011, When is a liability not a liability? textual analysis, dictionaries, and 10-Ks, *Journal of Finance* 66, 35–65.

Manela, Asaf, and Alan Moreira, 2017, News implied volatility and disaster concerns, *Journal of Financial Economics* 123, 137–162.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013, Efficient estimation of word representations in vector space, *arXiv.org* .

Newey, Whitney K., and Kenneth D. West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55.

Ramachandran, Lakshmi Shankar, and Jitendra Tayal, 2021, Mispricing, short-sale constraints, and the cross-section of option returns, *Journal of Financial Economics* 141, 297–321.

Sautner, Zacharias, Laurence Van Lent, Grigory Vilkov, and Ruishen Zhang, 2023, Firm-level climate change exposure, *Journal of Finance* 78, 1449–1498.

Schölkopf, Bernhard, and Alexander J. Smola, 2002, *Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond*, Adaptive Computation and Machine Learning (MIT Press, Cambridge, Mass).

Tetlock, Paul C., 2007, Giving content to investor sentiment: The role of media in the stock market, *Journal of Finance* 62, 1139–1168.

Tetlock, Paul C., 2010, Does public financial news resolve asymmetric information?, *Review of Financial Studies* 23, 3520–3557.

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More than words: Quantifying language to measure firms' fundamentals, *Journal of Finance* 63, 1437–1467.

Tian, Meng, and Liuren Wu, 2023, Limits of arbitrage and primary risk-taking in derivative securities, *Review of Asset Pricing Studies* 13, 405–439.

Vasquez, Aurelio, and Xiao Xiao, 2024, Default risk and option returns, *Management Science* 70, 2144–2167.

Vilkov, Grigory, 2023, Option-implied data and analysis.

Zhan, Xintong, Bing Han, Jie Cao, and Qing Tong, 2022, Option return predictability, *Review of Financial Studies* 35, 1394–1442.

Table 1
Summary Statistics

Table 1 presents the descriptive statistics of important variables used in this paper, alongside the correlations between textual predictors and quantitative determinants of option returns. The sample period is from January 1996 to November 2022. Panel A reports the time-series average of the cross-sectional summary statistics for several important variables. A delta-hedged call (put) option portfolio involves buying one contract of an equity call (put) and a short position of Δ shares of the underlying stock, where Δ is the Black-Scholes call (put) option delta. Delta-hedged option return is defined as the total dollar gain of the delta-hedged option portfolio scaled by the absolute value of the cost of the delta-hedged option portfolio at its formation date. *Call Option Return* (*Put Option Return*) is the return of the delta-hedged call (put) option portfolio with daily rebalancing. TP^{Call} (TP^{Put}) is the textual predictor extracted from news media for delta-hedged call (put) option returns using support vector regression model. *IVOL* is the idiosyncratic volatility computed as in Ang et al. (2006). *HV-IV* is the difference between realized volatility and implied volatility as in Goyal and Saretto (2009). *ILLIQ* is the natural logarithm of the stock illiquidity measure from Amihud (2002). *MFIS* (*MFIK*) is the model-free option-implied skewness (kurtosis), as in Bakshi and Kapadia (2003). *VOIV* is the volatility of the implied volatility, as in Cao et al. (2023). All variables in this table are at the monthly frequency, except for *HV-IV*, which is annualized. Panel B reports the time-series average of the cross-sectional Pearson correlations of textual predictors and various control variables in our study.

Panel A: Time-Series Average of Cross-sectional Summary Statistics for Important Variables

	Mean	Standard Deviation	10th Percentile	Lower Quartile	Median	Upper Quartile	90th Percentile
Call Option Return (%)	0.04	4.52	-3.33	-1.67	-0.29	1.13	3.20
Put Option Return (%)	-0.21	3.85	-3.18	-1.69	-0.42	0.90	2.77
TP^{Call} (%)	-0.33	0.22	-0.62	-0.47	-0.32	-0.18	-0.06
TP^{Put} (%)	-0.44	0.20	-0.71	-0.57	-0.44	-0.31	-0.20
IVOL (%)	1.93	1.37	0.86	1.12	1.58	2.33	3.36
HV-IV (%)	1.59	10.18	-7.39	-2.77	1.16	5.45	11.22
ILLIQ	-8.39	1.57	-10.31	-9.48	-8.53	-7.36	-6.26
MFIS	-0.49	0.44	-0.98	-0.70	-0.47	-0.26	-0.03
MFIK	4.40	1.36	3.26	3.52	3.99	4.86	6.11
VOIV (%)	5.91	4.23	3.17	3.91	5.00	6.59	8.92

Panel B: Time-series Average of Cross-sectional Correlations

	TP^{Put}	IVOL	HV-IV	ILLIQ	MFIS	MFIK	VOIV
TP^{Call}	0.78	-0.09	0.02	-0.09	-0.05	0.00	-0.04
TP^{Put}		-0.08	0.02	-0.09	-0.03	-0.01	-0.04
IVOL			0.11	0.36	0.19	-0.20	0.21
VRP				-0.06	-0.06	0.08	-0.03
ILLIQ					0.16	0.05	0.12
MFIS						-0.35	-0.03
MFIK							0.19

Table 2
Option Portfolios Sorted by Textual Predictors Using Support Vector Regression

Table 2 reports the average monthly excess returns to the delta-hedged option portfolios sorted by TP^{Call} (TP^{Put}). At the end of each month, we rank all underlying stocks into quintiles by their TP^{Call} (TP^{Put}). Detailed descriptions of TP^{Call} (TP^{Put}) are provided in Section 2.2. The portfolio is held for one month and rebalanced every day. This table reports the average return to the delta-hedged option portfolio for each quintile, as well as the (5 – 1) return spread (that is, the difference in returns between the portfolios of the highest and lowest quintiles). We also adjust the average returns using factor models and report the corresponding alphas. The first factor model is a seven-factor model includes the five stock factors in Fama and French (2015), the momentum factor, and the option factor in Coval and Shumway (2001). The second factor model is an option two-factor model includes an idiosyncratic volatility factor and a stock illiquidity factor as described in Zhan et al. (2022). The realization of the idiosyncratic volatility (illiquidity) factor is the (10-1) stock-value-weighted spread return for portfolios of daily-rebalanced delta-hedged option returns sorted on idiosyncratic volatility (natural logarithm of the Amihud stock illiquidity measure) of the underlying stock. We construct the option two-factor model for call option returns and put option returns, respectively. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey and West (1987) t statistics are reported in brackets. The symbols *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 – 1)**
Call	Average return	-0.25	0.00	0.03	0.09	0.31	0.56***
		(-2.09)	(0.02)	(0.23)	(0.74)	(2.11)	(6.64)
	7-Factor α	-0.18	0.13	0.09	0.15	0.38	0.56***
		(-1.12)	(0.76)	(0.58)	(1.00)	(1.99)	(5.78)
	Option 2-Factor α	-0.30	-0.06	-0.04	0.04	0.26	0.55***
		(-2.84)	(-0.51)	(-0.39)	(0.41)	(1.90)	(5.86)
		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 – 1)**
Put	Average return	-0.41	-0.22	-0.19	-0.15	-0.05	0.36***
		(-4.07)	(-2.02)	(-1.75)	(-1.40)	(-0.40)	(5.42)
	7-Factor α	-0.37	-0.16	-0.13	-0.10	0.01	0.38***
		(-2.90)	(-1.22)	(-0.93)	(-0.74)	(0.05)	(4.83)
	Option 2-Factor α	-0.43	-0.24	-0.20	-0.17	-0.05	0.38***
		(-3.90)	(-1.98)	(-1.67)	(-1.32)	(-0.34)	(4.71)

Table 3
Dependent Double Portfolio Sorting

In this table, we investigate whether several control variables can individually explain the predictability of SVR textual predictors using dependent double sorts. We first sort all options into tertiles based on a given control variable such as idiosyncratic volatility (*IVOL*), volatility deviation (*HV–IV*), Amihud stock illiquidity measure (*ILLIQ*), model-free implied skewness (*MFIS*), model-free implied kurtosis (*MFIK*), or volatility of implied volatility (*VOIV*). Then, within each tertile we further sort the options into quintiles based on the SVR textual predictors. Finally, we average returns for each textual predictor quintile across groups sorted by the control variable, yielding five control-variable adjusted quintile returns. Then we report the top-minus-bottom return spreads for the control-variable adjusted quintiles. We report the baseline results based on univariate sort (without control) in the first row, followed by the corresponding results after controlling for the variable labeled in each subsequent row. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey and West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Call Options	**Put Options**
Baseline	0.56***	0.36***
	(6.64)	(5.42)
IVOL	0.47***	0.28***
	(5.67)	(4.41)
HV-IV	0.52***	0.30***
	(6.14)	(4.41)
ILLIQ	0.47***	0.27***
	(5.45)	(4.10)
MFIS	0.54***	0.36***
	(6.41)	(5.58)
MFIK	0.55***	0.32***
	(6.20)	(5.07)
VOIV	0.53***	0.30***
	(6.37)	(4.64)

Table 4
Fama-MacBeth Regressions

This table reports the Fama-Macbeth regression results of the delta-hedged equity option returns on SVR textual predictors. Detailed descriptions of textual predictors and their constructions are provided in Section 2.2. The constructions of control variables are described in the Variable Definition. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey and West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Call		**Put**	
	(1)	**(2)**	**(3)**	**(4)**
TP	0.18***	0.12***	0.12***	0.08***
	(6.35)	(4.50)	(5.24)	(3.41)
IVOL		-0.25***		-0.25***
		(-5.97)		(-6.47)
HV-IV		0.25***		0.28***
		(5.55)		(7.61)
ILLIQ		0.05		0.03
		(1.15)		(1.01)
MFIS		-0.06**		0.18***
		(-2.19)		(6.56)
MFIK		0.05		0.15***
		(1.53)		(6.18)
VOIV		-0.08**		-0.05
		(-2.59)		(-1.53)
Adj.R^2 (%)	0.39	5.75	0.30	6.06
Obs	85,556	85,055	85,556	85,055

Table 5
Option Portfolios Sorted by Textual Predictors based on Different Machine-Learning Algorithms

Table 5 reports average monthly excess returns of the delta-hedged option portfolios sorted by textual predictors obtained by alternative machine learning algorithms. The rows labeled "SVR", "ENET", "RF", and "NN" reports portfolio sorting results by textual predictors extracted based on support vector regression, elastic net, random forest, and neural networks, respectively. Detailed descriptions of these predictors are provided in Section 3.2.1. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey and West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 − 1)**
Call	SVR	-0.25	0.00	0.03	0.09	0.31	0.56***
		(-2.09)	(0.02)	(0.23)	(0.74)	(2.11)	(6.64)
	ENET	-0.24	0.04	0.06	0.08	0.24	0.47***
		(-1.88)	(0.37)	(0.42)	(0.62)	(1.60)	(5.22)
	RF	-0.25	0.05	0.07	0.10	0.22	0.48***
		(-2.05)	(0.38)	(0.54)	(0.78)	(1.55)	(6.53)
	NN	-0.23	0.01	0.09	0.08	0.24	0.48***
		(-1.94)	(0.05)	(0.69)	(0.63)	(1.55)	(4.87)
		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 − 1)**
Put	SVR	-0.41	-0.22	-0.19	-0.15	-0.05	0.36***
		(-4.07)	(-2.02)	(-1.75)	(-1.40)	(-0.40)	(5.42)
	ENET	-0.43	-0.18	-0.20	-0.11	-0.09	0.33***
		(-4.25)	(-1.60)	(-1.89)	(-1.01)	(-0.86)	(5.69)
	RF	-0.40	-0.25	-0.20	-0.08	-0.09	0.31***
		(-4.05)	(-2.36)	(-1.83)	(-0.71)	(-0.74)	(4.45)
	NN	-0.39	-0.24	-0.16	-0.15	-0.08	0.31***
		(-3.94)	(-2.36)	(-1.43)	(-1.31)	(-0.67)	(5.02)

Table 6
Option Portfolios Sorted by Textual Predictors based on Alternative Feature Constructions

This table reports the average monthly excess returns of the delta-hedged option portfolios sorted by SVR textual predictors trained by using alternative word feature constructions. The rows labeled "Unigram", "Bigram", "Trigram", or "Unigram + Bigram" reports portfolio sorting results based on SVR textual predictors extracted based on different word features to train the model, including unigrams, bigrams, trigrams, and the combination of unigrams and bigrams, respectively. Detailed descriptions of these predictors are provided in Section 3.2.2. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey and West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 − 1)**
Call	Unigram	-0.25	0.00	0.03	0.09	0.31	0.56***
		(-2.09)	(0.02)	(0.23)	(0.74)	(2.11)	(6.64)
	Bigram	-0.28	0.01	0.05	0.15	0.25	0.53***
		(-2.21)	(0.07)	(0.39)	(1.18)	(1.73)	(6.50)
	Trigram	-0.26	-0.03	0.10	0.09	0.29	0.55***
		(-2.11)	(-0.26)	(0.75)	(0.71)	(1.97)	(6.54)
	Unigram + Bigram	-0.27	-0.02	0.04	0.13	0.30	0.57***
		(-2.21)	(-0.15)	(0.31)	(1.04)	(2.01)	(6.39)
		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 − 1)**
Put	Unigram	-0.41	-0.22	-0.19	-0.15	-0.05	0.36***
		(-4.07)	(-2.02)	(-1.75)	(-1.40)	(-0.40)	(5.42)
	Bigram	-0.39	-0.24	-0.19	-0.15	-0.04	0.35***
		(-3.82)	(-2.37)	(-1.81)	(-1.38)	(-0.34)	(5.24)
	Trigram	-0.40	-0.23	-0.22	-0.11	-0.05	0.36***
		(-4.05)	(-2.26)	(-2.27)	(-0.95)	(-0.36)	(5.44)
	Unigram + Bigram	-0.42	-0.20	-0.19	-0.18	-0.03	0.40***
		(-4.27)	(-1.81)	(-1.79)	(-1.66)	(-0.20)	(5.56)

Table 7
Option Portfolios Sorted by Textual Predictors based on Word Embeddings or Large Language Models

Table 7 reports average monthly excess returns of the delta-hedged option portfolios sorted by textual predictors obtained by implementing text representation techniques such as word embeddings and large language models. The rows labeled "TFIDF," "WE," "BERT," and "FinBERT" present the portfolio sorting results by textual predictors based on text representation methods: tf-idf vectorizer, word embeddings, BERT, and FinBERT, respectively. Detailed descriptions of these predictors are provided in Section 3.2.3. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey and West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 − 1)**
Call	TFIDF	-0.25	0.00	0.03	0.09	0.31	0.56***
		(-2.09)	(0.02)	(0.23)	(0.74)	(2.11)	(6.64)
	WE	-0.24	-0.03	0.08	0.11	0.26	0.50***
		(-1.85)	(-0.28)	(0.61)	(0.89)	(1.79)	(6.17)
	BERT	-0.28	-0.00	0.03	0.10	0.33	0.61***
		(-2.28)	(-0.02)	(0.23)	(0.79)	(2.28)	(7.44)
	FinBERT	-0.24	-0.04	0.05	0.13	0.28	0.52***
		(-1.99)	(-0.31)	(0.38)	(1.04)	(1.86)	(5.70)
		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 − 1)**
Put	TFIDF	-0.41	-0.22	-0.19	-0.15	-0.05	0.36***
		(-4.07)	(-2.02)	(-1.75)	(-1.40)	(-0.40)	(5.42)
	WE	-0.38	-0.26	-0.19	-0.16	-0.03	0.36***
		(-3.78)	(-2.69)	(-1.78)	(-1.47)	(-0.20)	(4.35)
	BERT	-0.39	-0.25	-0.25	-0.15	0.03	0.42***
		(-3.65)	(-2.52)	(-2.44)	(-1.40)	(0.24)	(5.43)
	FinBERT	-0.39	-0.25	-0.18	-0.16	-0.04	0.35***
		(-3.95)	(-2.28)	(-1.69)	(-1.48)	(-0.30)	(5.27)

Table 8
Option Portfolios Sorted by Textual Predictors in Different Market Conditions

Table 8 reports the average monthly excess returns of the delta-hedged option portfolios sorted by ML textual predictors in different market conditions. The sentiment index is constructed in Baker and Wurgler (2006). Volatility is measured using CBOE VIX index. We use the equal-weighted Amihud (2002) stock illiquidity measure of individual stocks as a proxy for the market-level liquidity. "High Sentiment" ("Low Sentiment") is defined as periods during which the sentiment index is higher (lower) than the median sentiment index for the entire sample period. "High Volatility" ("Low Volatility") is defined as periods during which the VIX index is higher (lower) than the median VIX index for the entire sample period. "High Liquidity" ("Low Liquidity") is defined as periods during which the market-level illiquidity measure is lower (higher) than the median market-level illiquidity measure for the entire sample period. Recession periods are identified based on the recession timelines provided by the National Bureau of Economic Research (NBER).To adjust for serial correlations, robust Newey and West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

	1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 − 1)**
Panel A: Subperiod Analysis for Call Options						
High Sentiment	-0.02	0.26	0.32	0.34	0.62	0.64***
	(-0.10)	(1.16)	(1.62)	(1.63)	(2.87)	(5.84)
Low Sentiment	-0.49	-0.25	-0.26	-0.15	0.01	0.49***
	(-4.04)	(-1.88)	(-2.06)	(-1.14)	(0.04)	(4.24)
High Volatility	-0.25	-0.01	0.05	0.12	0.41	0.65***
	(-1.11)	(-0.04)	(0.21)	(0.53)	(1.48)	(4.23)
Low Volatility	-0.26	0.01	0.01	0.07	0.22	0.48***
	(-3.28)	(0.16)	(0.14)	(0.79)	(2.35)	(6.29)
High Liquidity	-0.22	0.08	0.03	0.08	0.25	0.47***
	(-1.66)	(0.50)	(0.21)	(0.53)	(1.75)	(6.49)
Low Liquidity	-0.29	-0.10	0.03	0.11	0.40	0.69***
	(-1.42)	(-0.43)	(0.13)	(0.52)	(1.40)	(4.05)
NBER Expansion	-0.23	-0.01	0.04	0.10	0.25	0.48***
	(-1.91)	(-0.11)	(0.33)	(0.76)	(1.95)	(6.89)
NBER Recession	-0.41	0.14	-0.07	0.06	0.85	1.25***
	(-0.90)	(0.29)	(-0.14)	(0.11)	(1.01)	(2.83)
Panel B: Subperiod Analysis for Put Options						
High Sentiment	-0.29	-0.07	0.03	-0.00	0.19	0.48***
	(-1.72)	(-0.45)	(0.16)	(-0.00)	(0.99)	(4.80)
Low Sentiment	-0.53	-0.37	-0.41	-0.30	-0.29	0.25***
	(-4.99)	(-2.52)	(-3.55)	(-2.66)	(-2.29)	(3.09)
High Volatility	-0.42	-0.24	-0.20	-0.12	0.03	0.45***
	(-2.44)	(-1.37)	(-1.10)	(-0.68)	(0.12)	(4.12)
Low Volatility	-0.40	-0.20	-0.17	-0.17	-0.12	0.28***
	(-4.51)	(-1.70)	(-1.95)	(-2.15)	(-1.23)	(3.67)
High Liquidity	-0.33	-0.11	-0.08	-0.09	-0.03	0.31***
	(-2.76)	(-0.85)	(-0.66)	(-0.78)	(-0.22)	(4.06)
Low Liquidity	-0.51	-0.36	-0.33	-0.22	-0.08	0.44***
	(-3.04)	(-2.03)	(-1.77)	(-1.17)	(-0.34)	(3.74)
NBER Expansion	-0.40	-0.21	-0.16	-0.16	-0.09	0.31***
	(-3.93)	(-1.91)	(-1.55)	(-1.66)	(-0.89)	(5.48)
NBER Recession	-0.49	-0.31	-0.39	-0.05	0.30	0.79**
	(-1.29)	(-0.75)	(-0.81)	(-0.08)	(0.47)	(2.36)

Table 9
Predictability for Returns of Alternative Portfolios

Table 9 presents the predictability of textual predictors for the returns of vanilla straddle portfolios, delta-neutral straddle portfolios, zero-beta straddle portfolios, and skewness portfolios. A vanilla straddle portfolio is constructed by taking long positions in at-the-money call and put options written on the same underlying stock with the same moneyness and maturity. Delta-neutral straddle portfolios and zero-beta straddle portfolios are created by taking long positions and assigning different weights to at-the-money call and put options. A skewness portfolio is constructed by taking long positions in out-of-the-money call and put options, as described in Bali and Murray (2013). Detailed constructions of these portfolios can be found in Section 3.2.4. The textual predictors are obtained from the SVR model with returns of these portfolios as the target variables. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey and West (1987) t-statistics are reported in brackets. The symbols *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

	1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 – 1)**
Vanilla Straddle	-3.29	-2.51	-1.94	-1.97	-1.12	2.17***
	(-3.87)	(-2.75)	(-2.04)	(-2.18)	(-1.10)	(3.12)
Delta-neutral Straddle	-4.88	-2.64	-2.67	-1.70	-0.40	4.48***
	(-4.04)	(-2.23)	(-2.13)	(-1.34)	(-0.32)	(5.38)
Zero-beta Straddle	-3.93	-3.05	-2.68	-2.04	-1.14	2.78***
	(-3.85)	(-3.02)	(-2.65)	(-1.93)	(-0.99)	(4.17)
Skewness Portfolio	-0.16	0.09	0.05	0.10	0.20	0.36***
	(-1.33)	(0.68)	(0.41)	(0.74)	(1.61)	(3.40)

Table 10
Overlap Analysis of Dictionaries

Table 10 reports overlap scores for various dictionaries for delta-hedged option return determinants. *Call* (*Put*) refers to the option return dictionaries for call (put) option returns. The dictionaries associated with option return determinants are named accordingly. The option return determinants are: ΔIV is the percentage change of implied volatility from month t to $t+1$. *IVOL* is the idiosyncratic volatility computed as in Ang et al. (2006). *HV-IV* is the difference between realized volatility and implied volatility at time t. *ILLIQ* is the Amihud stock illiquidity measure as in Amihud (2002). *VOIV* is the volatility of the implied volatility, calculated as the standard deviation of implied volatility during month t. *MFIS* (*MFIK*) is the model-free option-implied skewness (kurtosis), as in Bakshi and Kapadia (2003), inferred from a cross section of out of the money calls and puts at the end of the time t. For each dictionary, the overlap score is given by:

$$Overlap\ Score = \frac{C - W}{C + N + W}$$

where C is the number of words that overlap correctly, N is the number of non-overlapping words, and W is the number of words that overlap incorrectly. Section 4.2.1 explains in detail how a word is classified as correctly or incorrectly overlapped with the option return dictionary. The overlap score ranges from -1 to 1, where a higher value indicates a higher degree of similarity between the information sets of the given dictionary and the dictionary for the delta-hedged option returns. The logic of this calculation can be found in Section 4.2.1.

	Call (%)	**Put (%)**
ΔIV	21.80	21.00
IVOL	20.90	19.70
MFIS	17.71	13.03
MFIK	10.60	8.50
VOIV	11.35	10.35
ILLIQ	8.80	10.45
HV-IV	-0.35	-0.55

Table 11
Decomposition of Textual Predictors

Table 11 reports the results of the regression analysis that decomposes the SVR textual predictors by projected values of various option return determinants. The projected values are obtained by applying the SVR algorithm to each option return determinant using the news data as the information set. The dependent variable is the SVR predicted option returns and the independent variables are the projected values of each option return determinant. The definition of option return determinants is the same as those in Table 10. Independent variables are standardized to have zero mean and unit standard deviation. The coefficients in this table are multiplied by 100 for presentation. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey and West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Decomposition of TP^{Call}								
ΔIV	0.07***							0.09***
	(8.20)							(13.01)
IVOL		-0.06***						-0.05***
		(-5.94)						(-5.12)
ILLIQ			-0.04***					0.01
			(-4.74)					(1.36)
MFIS				-0.07***				-0.07***
				(-6.64)				(-5.64)
MFIK					0.04***			-0.05***
					(5.29)			(-5.62)
VOIV						-0.03***		-0.02***
						(-3.79)		(-4.61)
HV-IV							-0.01	-0.02***
							(-1.63)	(-3.65)
Adj.R^2 (%)	15.48	13.59	8.43	13.97	7.39	6.94	6.06	43.44
Obs	85,055	85,055	85,055	85,055	85,055	85,055	85,055	85,055
Panel B: Decomposition of TP^{Put}								
ΔIV	0.06***							0.08***
	(8.28)							(11.87)
IVOL		-0.04***						-0.03***
		(-6.41)						(-4.40)
ILLIQ			-0.04***					-0.01
			(-4.78)					(-1.19)
MFIS				-0.04***				-0.04***
				(-5.62)				(-3.76)
MFIK					0.02***			-0.03***
					(3.52)			(-4.97)
VOIV						-0.03***		-0.02***
						(-5.04)		(-5.49)
HV-IV							-0.01	-0.02***
							(-1.36)	(-3.50)
Adj.R^2 (%)	13.48	10.71	7.73	10.92	5.30	6.03	4.53	36.86
Obs	85,055	85,055	85,055	85,055	85,055	85,055	85,055	85,055

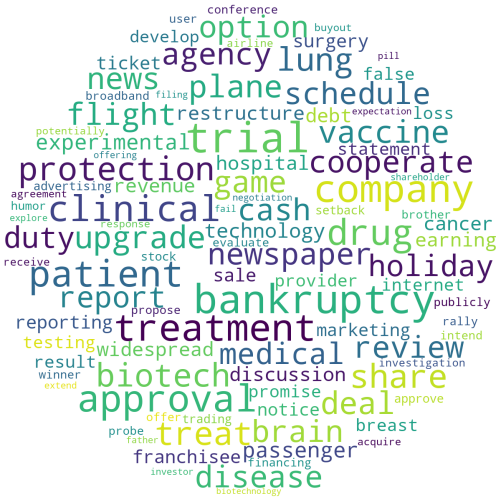
Table 12
Generalizability to Earnings Conference Call Transcripts

Table 12 presents the panel regression results of delta-hedged option returns on textual predictors obtained by applying the SVR model trained on the news media corpus to the earnings conference call transcripts. The left-hand side variable is the delta-hedged option return of firm i in month $t+1$ if firm i have a conference call in month t. The definition and construction of TP^{Conf} can be found in Section 4.2.3. Detailed constructions of control variables can be found in the Variable Definition. We include the time fixed effect and cluster standard errors by time. The sample period is from November 2005 to December 2021. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Call		**Put**	
	(1)	**(2)**	**(3)**	**(4)**
TP^{Conf}	0.567***	0.377***	0.263***	0.161**
	(5.75)	(4.03)	(3.72)	(2.48)
IVOL		-0.274***		-0.257***
		(-4.84)		(-6.53)
HV-IV		0.418***		0.306***
		(4.29)		(5.40)
ILLIQ		-0.080*		-0.024
		(-1.76)		(-0.71)
MFIS		-0.075***		0.061**
		(-3.15)		(2.48)
MFIK		0.162***		0.192***
		(5.52)		(6.70)
VOIV		0.092**		0.109***
		(2.51)		(3.38)
Time FE	Yes	Yes	Yes	Yes
Adj.R^2 (%)	8.30	9.21	7.31	8.17
Obs	55,881	55,399	55,881	55,399

Figure 1. Word Cloud of Option Return Dictionaries

(a) Call Positive



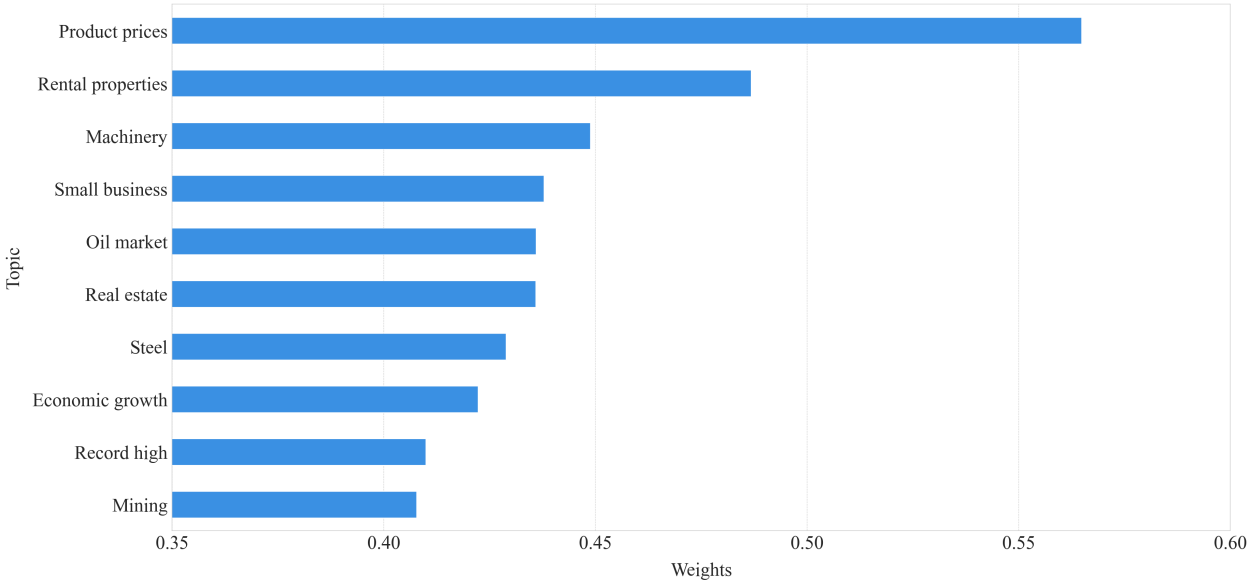
(b) Call Negative

(c) Put Positive

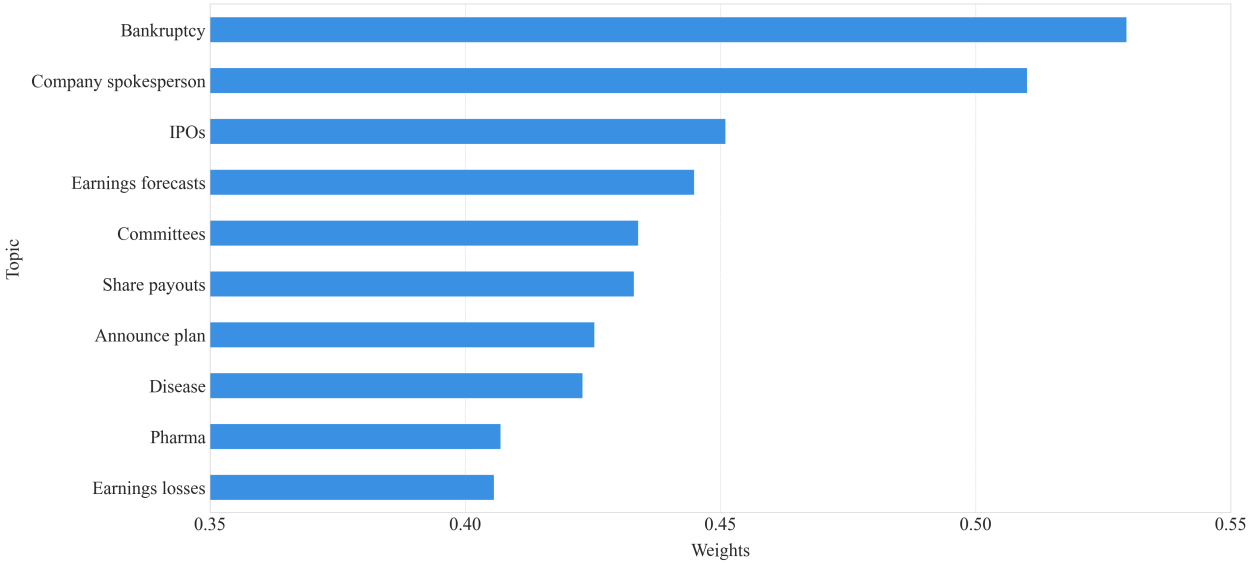
(d) Put Negative

Figure 1: This figure reports the top 100 words in each option return dictionary. Font size of a word is proportional to its time of appearance as positive or negative related to equity option returns.

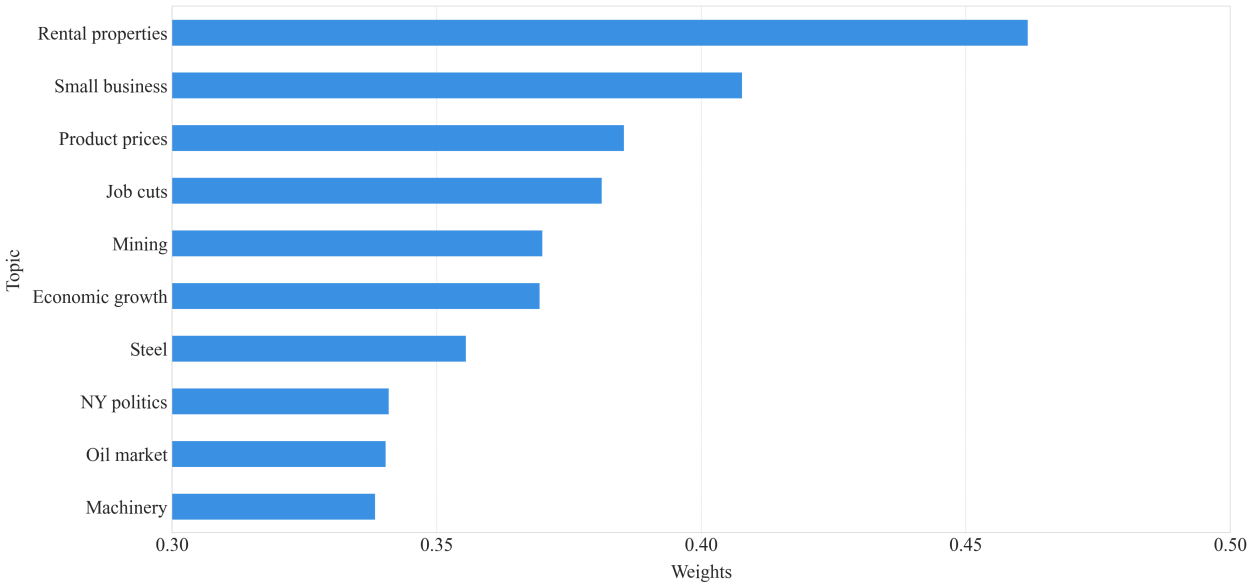
Figure 2. Top 10 Topics for Option Return Dictionary



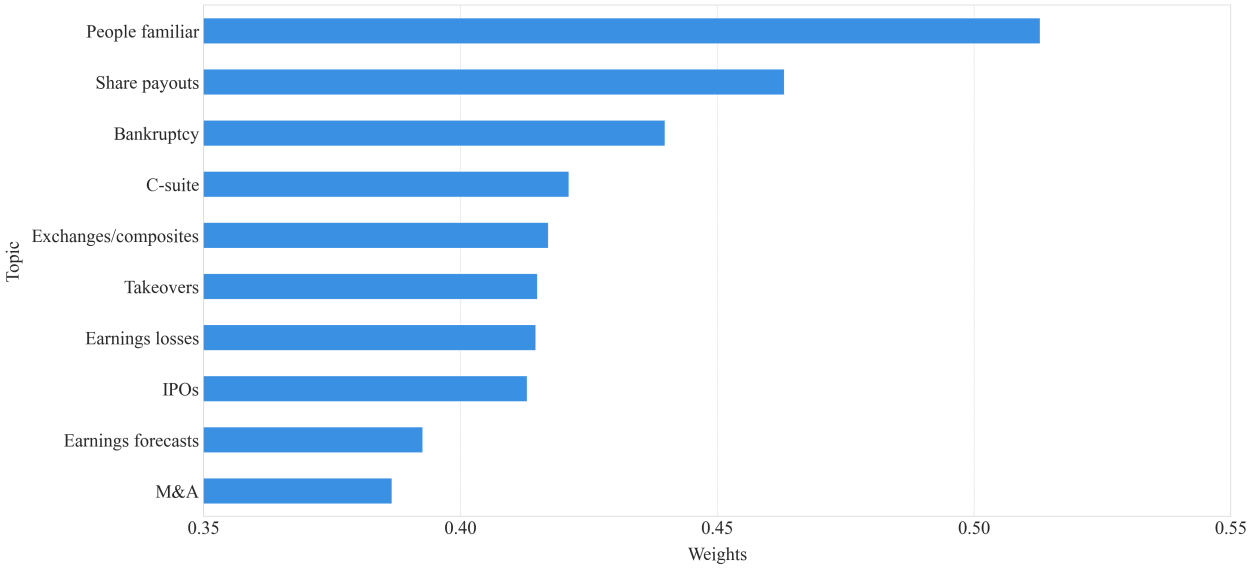
(a) Call Positive



(b) Call Negative



(c) Put Positive



(d) Put Negative

Figure 2: This figure displays the top 10 topics identified in each option return dictionary. We apply the word weights by topics derived from Bybee et al. (2021) to obtain the relative importance of topics within our option return dictionaries. Specifically, for each term in each option return dictionary, we get its vector of weights across the 180 topics. We then aggregate the word weights at the topic level by summing the weights across words in the given dictionary. We plot the top 10 topics based on their total weights for each option return dictionary.

Figure 3. Causal Diagram for the Predictability of Textual Information

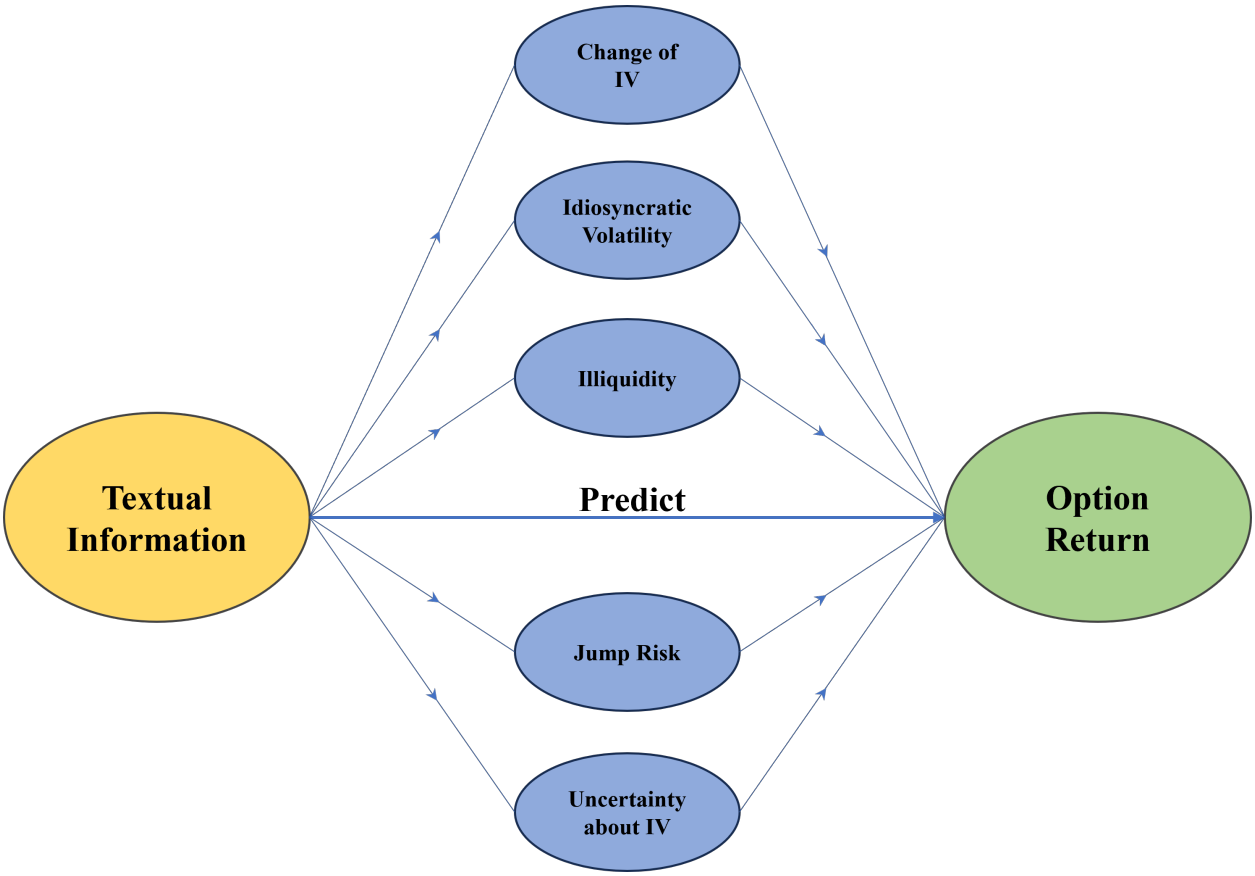


Figure 3: This figure illustrates the underlying logic of our economic mechanism analysis.

Figure 4. Illustration of the Calculation of Overlap Score

		Call Option		
		Positive	Non-Overlap	Negative
ΔIV	Positive	TRUE (359)	510	FALSE (131)
	Negative	FALSE (134)	524	TRUE (342)

(a) Call

		Put Option		
		Positive	Non-Overlap	Negative
ΔIV	Positive	TRUE (349)	514	FALSE (137)
	Negative	FALSE (129)	534	TRUE (337)

(b) Put

Figure 4: This figure shows how the overlap score is calculated between dictionaries for change of implied volatility and call option returns. $C = (359+342 = 701), N = (510+524 = 1034), W = (134+131 = 265)$, so that the overlap score is $(701-265)/(701+1034+265) = 0.218$. The calculation of overlap score of dictionaries for change of implied volatility and put option returns is similar.

Appendix for Forecasting Option Returns with News

Variable Definition

	Textual Predictors
TP	The textual predictors extracted from news media using the support vector regression model. TP are estimated separately for call options (TP^{Call}) and put options (TP^{Put}).
	Control Variables
IVOL	Idiosyncratic volatility, defined as the standard deviation of daily return residuals from regressions of daily returns on the Fama-French 3-factor model over the previous month, following Ang et al. (2006). We require at least 15 observations for the regression.
HV-IV	Volatility deviation, defined as the difference between realized volatility and implied volatility following Goyal and Saretto (2009). Realized volatility is the standard deviation of daily realized stock returns over the past year. Implied volatility is the average of ATM call and put implied volatility with 30-day maturity, obtained from the Volatility Surface dataset of OptionMetrics IvyDB database.
VOIV	Volatility of the implied volatility, calculated as the standard deviation of the daily percentage change of option implied volatility over the trading days within a given month, and the implied volatility is the average of at-the-money implied volatility of call and put options obtained from Volatility Surface provided by OptionMetrics IvyDB database.
ILLIQ	Amihud illiquidity measure Amihud (2002), calculated as dividing the absolute daily return of a stock by its dollar volume over the past one month.
MFIS (MFIK)	Model-free option-implied skewness (kurtosis), as in Bakshi and Kapadia (2003), inferred from a cross section of out of the money calls and puts at the end of the last month. We thank Grigory Vilkov (Vilkov (2023)) for providing the Python code to calculate these measures, and the corresponding code and data can be found via https://www.vilkov.net/codedata.html

Other Variables	
ΔIV	The future implied volatility changes over the next month.
LM_Sentiment	Dictionary-based sentiment measure derived from the LM dictionary. For a given firm, LM_Sentiment is defined as the difference between the positive and negative words detected based on the LM dictionary in the aggregated article for each month, scaled by the length of the aggregated article. LM dictionary is the Loughran-McDonald dictionary from Loughran and McDonald (2011).
LM_Uncertainty	Dictionary-based uncertainty measure derived from the LM dictionary. For a given firm, LM_Uncertainty is defined as the number of uncertainty words detected based on the LM dictionary in the aggregated article for each month, scaled by the length of the aggregated article. LM dictionary is the Loughran-McDonald dictionary from Loughran and McDonald (2011).

Table A1
Sample Coverage of Underlying Stocks

Table A1 provides details about the stock-month sample for the underlying stocks covered in our analysis. Panel A reports the time-series summary statistics of our sample coverage, and Panel B reports the time-series average of cross-sectional distributions. Panel C reports the time-series average of the Fama-French 12-industry distribution for the sample of stocks covered in our analysis and full CRSP sample. The percent coverage of stock universe (EW) is the number of sample stocks, divided by the total number of CRSP stocks. The percent coverage of the stock universe (VW) is the total market capitalization of sample stocks divided by the total market value of all CRSP stocks. Optionable stocks are defined as stocks with valid options at the end of each month. Firm size is the firm's market capitalization. Book-to-market is the fiscal year-end book value of common equity divided by the calendar year-end market value of equity. Institutional ownership is the percentage of common stocks owned by institutions in the previous quarter. Analyst coverage is the number of analysts following the firm in the previous month. The sample period is from January 1996 to November 2022.

Panel A: Time-Series Distribution (323 Monthly Obs.)

January 1996–November 2022	Mean	Standard Deviation	10th Percentile	Lower Quartile	Median	Upper Quartile	90th Percentile
Stock % coverage of stock universe (EW)	3.69	1.11	1.70	3.17	3.94	4.47	4.95
Stock % coverage of stock universe (VW)	33.55	8.06	22.95	27.92	33.22	38.72	44.93
Stock % coverage of optionable stocks (EW)	9.54	4.31	3.09	6.32	9.73	13.44	14.98
Stock % coverage of optionable stocks (VW)	36.10	9.40	23.65	29.15	35.59	42.53	49.44
Stock % traded at NYSE/AMEX	69.62	5.08	61.18	67.13	70.71	73.05	75.27
Stock % included in S&P500 index	60.60	7.77	54.13	56.62	59.51	62.73	66.91
Stock % already included in last month	48.08	5.57	40.00	44.68	48.64	52.05	54.46

Panel B: Time-Series Average of Cross-Sectional Distributions (89,895 Stock-Month Obs.)

January 1996–November	Mean	Standard Deviation	10th Percentile	Lower Quartile	Median	Upper Quartile	90th Percentile
Firm size in billions	33.16	67.65	1.37	3.35	10.78	31.90	83.52
Firm size CRSP percentile (%)	86.99	10.69	71.47	82.63	90.96	94.75	96.18
Firm book-to-market CRSP percentile (%)	34.49	25.03	5.16	13.25	29.54	52.92	72.70
Institutional Ownership (%)	65.67	16.89	42.10	57.60	68.98	77.73	83.69
Analyst Coverage	12.36	6.58	4.65	7.42	11.46	16.18	21.39

Panel C: Time-Series Average of Industry Distribution (%)

FF-12 Industry	Paper Sample	CRSP Sample	FF-12 Industry	Paper Sample	CRSP Sample
Consumer nondurables	8.24	4.70	Telecom	2.78	2.66
Consumer durables	2.83	2.21	Utilities	2.56	2.25
Manufacturing	9.67	8.99	Wholesale	13.93	8.59
Energy	3.37	3.76	Healthcare	7.65	12.48
Chemicals	3.28	2.06	Finance	14.32	20.89
Business Equipment	19.54	17.76	Others	11.90	13.64

Table A2
N-gram Coverage

Table A2 reports the average percentage coverage of token frequencies in the corpus that are covered by the top N most frequent tokens. In each training iteration, we divide the token frequency covered by the most frequent N tokens by the total token frequency of the training sample corpus to calculate the percentage coverage of the top N tokens. After completing all training iterations, we compute the time-series average of the percentage coverage.

N	Unigram (%)	Bigram (%)	Trigram (%)
1,000	64.29	4.96	1.29
2,000	78.10	6.85	1.71
3,000	84.67	8.22	2.01
4,000	88.51	9.34	2.26
5,000	91.02	10.29	2.48
6,000	92.80	11.13	2.68
7,000	94.10	11.89	2.86
8,000	95.10	12.59	3.03
9,000	95.89	13.23	3.19
10,000	96.52	13.83	3.34
40,000	99.91	24.32	6.47
80,000	100.00	31.74	9.29

Table A3
Fama-MacBeth Regressions Controlling for Dictionary-Based Measures

This table reports the Fama-Macbeth regression results of the delta-hedged equity option returns on SVR textual predictors with dictionary-based measures as additional control variables. LM_Sentiment and LM_Uncertainty are two dictionary-based measures derived from the Loughran and McDonald (2011) dictionary, and the constructions of them can be found in the Variable Definition. Detailed descriptions of textual predictors and their constructions are provided in Section 2.2. The constructions of control variables are described in the Variable Definition. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey and West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Call			**Put**		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
SVR	0.18***	0.18***	0.11***	0.12***	0.12***	0.08***
	(6.18)	(6.24)	(4.22)	(5.10)	(5.12)	(3.29)
LM_Sentiment	0.03*		0.03	0.02		0.01
	(1.67)		(1.46)	(1.06)		(0.76)
LM_Uncertainty		0.01	0.03		-0.03*	-0.02
		(0.71)	(1.37)		(-1.85)	(-1.18)
IVOL			-0.24***			-0.25***
			(-5.82)			(-6.49)
VRP			0.25***			0.28***
			(5.59)			(7.71)
ILLIQ			0.05			0.03
			(1.14)			(0.95)
MFIS			-0.06**			0.18***
			(-2.19)			(6.59)
MFIK			0.04			0.15***
			(1.45)			(6.15)
VOIV			-0.08***			-0.05
			(-2.71)			(-1.50)
Adj.R^2 (%)	0.53	0.53	5.99	0.43	0.44	6.26
Obs	85,546	85,546	85,045	85,546	85,546	85,045

Table A4
Option Portfolios Sorted by SVR Textual Predictors with Different Numbers of Word Features

Table A4 presents the average monthly excess returns of delta-hedged option portfolios sorted by textual predictors developed using SVR but with different numbers of word features. The rows labeled "Baseline", "4,000," "6,000," and "8,000" correspond to the portfolio sorting results using textual predictors derived from the tf-idf matrix comprising the most frequent 10,000, 4,000, 6,000, and 8,000 words, respectively. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey and West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 − 1)**
Call	Baseline	-0.25	0.00	0.03	0.09	0.31	0.56***
		(-2.09)	(0.02)	(0.23)	(0.74)	(2.11)	(6.64)
	4,000	-0.23	0.01	0.06	0.09	0.26	0.49***
		(-1.93)	(0.11)	(0.45)	(0.66)	(1.78)	(5.95)
	6,000	-0.26	0.00	0.04	0.14	0.27	0.53***
		(-2.11)	(0.01)	(0.33)	(1.07)	(1.79)	(6.23)
	8,000	-0.26	0.01	0.06	0.10	0.27	0.53***
		(-2.12)	(0.07)	(0.45)	(0.77)	(1.92)	(6.38)
		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 − 1)**
Put	Baseline	-0.41	-0.22	-0.19	-0.15	-0.05	0.36***
		(-4.07)	(-2.02)	(-1.75)	(-1.40)	(-0.40)	(5.42)
	4,000	-0.36	-0.23	-0.20	-0.17	-0.06	0.30***
		(-3.33)	(-2.25)	(-1.82)	(-1.58)	(-0.52)	(3.84)
	6,000	-0.40	-0.21	-0.18	-0.16	-0.07	0.33***
		(-4.00)	(-1.85)	(-1.73)	(-1.40)	(-0.62)	(5.25)
	8,000	-0.41	-0.23	-0.16	-0.14	-0.07	0.34***
		(-4.24)	(-2.05)	(-1.58)	(-1.22)	(-0.60)	(5.04)

Table A5
Option Portfolios Sorted by SVR Textual Predictors Across Different Rolling Windows

Table A5 presents the average monthly excess returns of delta-hedged option portfolios sorted by textual predictors obtained by SVR over various rolling window periods. The rows labeled "Baseline", "3-month", "9-month", and "12-month" correspond to the portfolio sorting results when rolling window is 6 months, 3 months, 9 months, and 12 months, respectively. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey and West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 − 1)**
Call	Baseline	-0.25	0.00	0.03	0.09	0.31	0.56***
		(-2.09)	(0.02)	(0.23)	(0.74)	(2.11)	(6.64)
	3-month	-0.23	-0.03	0.06	0.12	0.27	0.51***
		(-1.91)	(-0.24)	(0.44)	(0.97)	(1.88)	(5.99)
	9-month	-0.26	0.05	0.07	0.08	0.23	0.48***
		(-2.03)	(0.44)	(0.45)	(0.71)	(1.57)	(6.47)
	12-month	-0.24	-0.01	0.07	0.13	0.23	0.47***
		(-2.01)	(-0.06)	(0.54)	(1.05)	(1.49)	(5.71)
		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 − 1)**
Put	Baseline	-0.41	-0.22	-0.19	-0.15	-0.05	0.36***
		(-4.07)	(-2.02)	(-1.75)	(-1.40)	(-0.40)	(5.42)
	3-month	-0.39	-0.23	-0.16	-0.16	-0.07	0.32***
		(-3.89)	(-1.92)	(-1.43)	(-1.64)	(-0.52)	(4.52)
	9-month	-0.39	-0.21	-0.18	-0.15	-0.08	0.31***
		(-3.75)	(-1.86)	(-1.74)	(-1.36)	(-0.66)	(5.37)
	12-month	-0.36	-0.26	-0.22	-0.13	-0.03	0.33***
		(-3.38)	(-2.44)	(-2.13)	(-1.14)	(-0.28)	(4.23)

Table A6
Correlations among Textual Predictors based on Alternative Machine-learning Algorithms or Text Representation Techniques

Table A6 reports the time-series average of the cross-sectional correlations among textual predictors based on different machine-learning algorithms. "SVR"/"ENET"/"RF"/"NN" represents textual predictors extracted based on support vector regression, elastic net, random forest, neural networks. "WE"/"BERT"/"FinBERT" represents textual predictors when the text data is tokenized using Word2Vec, BERT, and FinBERT, respectively.

		ENET	RF	NN	WE	BERT	FinBERT
Call	SVR	0.70	0.48	0.66	0.61	0.54	0.53
	ENET		0.50	0.82	0.52	0.48	0.46
	RF			0.51	0.40	0.36	0.35
	NN				0.51	0.49	0.47
	WE					0.59	0.57
	BERT						0.81
		ENET	RF	NN	WE	BERT	FinBERT
Put	SVR	0.56	0.49	0.59	0.51	0.42	0.44
	ENET		0.47	0.74	0.40	0.38	0.39
	RF			0.54	0.34	0.33	0.33
	NN				0.42	0.39	0.43
	WE					0.42	0.44
	BERT						0.66

Table A7
Option Return Dictionaries with Unigrams

This table lists the top 100 terms in each option return dictionary. For abbreviation, we list the first 100 terms with the highest frequency of positive or negative occurrences. The sample period is from January 1996 to November 2022.

	Positive	Negative
Call	bank, energy, article, business, field, land, manager, region, price, barrel, sanction, large, average, cycle, packaging, overall, open, grow, storm, operation, invest, ensure, need, remain, metal, economic, growth, global, chain, electronic, railroad, rate, brokerage, capability, score, plant, stand, strong, chemical, lead, license, pull, area, lift, residential, capital, range, make, insurance, state, producer, woman, management, industrial, controversy, view, dividend, major, single, city, begin, gold, architect, emerge, aerospace, ground, production, branch, start, impact, division, mine, wood, compete, portfolio, wage, size, coal, supply, flood, piece, employee, secret, hold, commodity, investment, corporate, wine, transport, host, center, expand, line, semiconductor, capacity, inflation, world, senior, customer, pump	trial, bankruptcy, company, approval, treatment, patient, clinical, drug, treat, protection, share, biotech, lung, deal, upgrade, vaccine, plane, agency, newspaper, option, cash, flight, review, news, brain, game, cooperate, disease, schedule, holiday, report, duty, medical, experimental, passenger, technology, revenue, earning, widespread, discussion, cancer, hospital, debt, restructure, reporting, franchisee, surgery, sale, ticket, marketing, notice, false, statement, provider, internet, promise, result, breast, testing, loss, develop, advertising, stock, trading, setback, approve, financing, conference, rally, humor, probe, broadband, winner, evaluate, user, publicly, intend, offer, propose, brother, investigation, receive, airline, potentially, agreement, buyout, acquire, response, expectation, fail, explore, filing, extend, shareholder, offering, pill, investor, father, negotiation, biotechnology
Put	need, dollar, manager, article, average, open, worker, field, area, region, driver, global, plant, remain, chain, processor, steel, large, score, sanction, author, barrel, build, estate, central, acquisition, prescription, fiber, rate, state, mine, speed, paper, grow, impact, producer, strong, glass, problem, view, overall, leader, main, resource, chemical, building, chip, laptop, policy, train, register, pharmacy, dividend, stake, sponsor, location, brokerage, selling, presence, limit, computer, gain, reach, trust, good, corporate, license, price, urban, thank, high, innovation, card, increase, expand, compete, major, ground, controversy, proprietary, telecommunications, maker, employer, growth, drugstore, business, energy, golf, computing, host, putt, ruling, employee, house, ship, young, knock, format, class, emerge	company, bankruptcy, review, financing, sale, jump, debt, trading, schedule, short, stock, vaccine, share, protection, lung, internet, buyout, trial, credit, result, drug, approval, technology, offering, expectation, analyst, seller, alternative, restructure, rebound, cigarette, restructuring, publish, lender, false, bondholder, response, tobacco, interested, temporarily, shareholder, marketing, lower, deal, profit, cancer, tracking, subscriber, loss, intend, clinical, earning, satellite, test, treatment, cell, familiar, guarantee, utility, symbol, experimental, filing, treat, gift, offer, sell, statement, evaluate, divestiture, infection, tracker, potentially, franchisee, rally, slot, relate, agency, halt, holiday, chance, biotech, newspaper, investigate, booking, content, buyer, investor, initial, online, available, potential, shoe, revenue, casino, report, patient, distribution, machine, traveler, investigation

Table A8
Option Return Dictionaries with Bigrams

This table lists the important bigram tokens in each option return dictionary. For abbreviation, we list the first 100 tokens with the highest frequency of positive or negative occurrences. The sample period is from January 1996 to November 2022.

	Positive	Negative
Call	real estate, interest rate, asset management, mutual fund, supply chain, percentage point, senior vice, world large, credit card, increase number, emerge market, young people, financial service, company offer, high profile, energy price, raise money, service company, double digit, insurance company, high yield, people live, spend time, black white, want know, work hard, central bank, vice chairman, earning growth, service business, general manager, profit margin, group company, investment bank, work force, people know, attorney general, license article, fast grow, tell investor, pension scheme, pension fund, tell story, firm base, company fund, equipment maker, take place, business development, office building, small business, venture capital, investment banking, plan launch, profit forecast, large bank, consumer product, investment officer, insurance industry, company need, high rate, employ people, class action, chief investment, large home, expect generate, include work, economic growth, federal regulator, fund invest, company provide, high rise, sell business, operating system, financial market, company move, record high, human resource, think people, growth come, office space, investment vehicle, expect cost, live work, blue chip, company value, bond fund, chip company, phone service, company come, presidential campaign, heart disease, large company, people come, public relation, chief economist, home improvement, number people, fund fund, rest world, national security	clinical trial, share rise, company share, drug company, patient receive, drug development, late stage, report earning, bankruptcy protection, company stock, drug market, expect earning, develop drug, company report, profit share, share share, earning share, share analyst, deal value, generic drug, share accord, share revenue, drug approve, accord people, share period, cancer drug, composite trading, earning report, file bankruptcy, stock price, income rise, price share, approve drug, share jump, share exclude, company plan, gain share, total company, income share, experimental drug, treat patient, stock rise, person familiar, revenue fall, drug maker, cancer treatment, regulatory approval, statement company, fall company, familiar matter, cash stock, post loss, report revenue, company announce, share company, drug call, revenue earning, market close, lose value, accord analyst, emerge bankruptcy, regulatory filing, biotech company, accord person, group investor, breast cancer, patent infringement, share plunge, company record, flight attendant, compare share, analyst expect, company sell, general counsel, medical device, agree acquire, bankruptcy court, expect share, trading share, save money, expect report, rise trading, restructuring plan, send share, home buyer, suit file, stock deal, increase revenue, expect loss, expect revenue, time charge, frequent flier, sale fall, estimate analyst, public health, medium company, ovarian cancer, share drop, company statement, share trading

Put	real estate, credit card, small business, shopping center, asset management, high price, vice president, license article, young people, operating system, interest rate, general manager, mobile phone, joint venture, increase number, equipment maker, insurance company, growth come, senior vice, health insurer, father retire, supply chain, class action, financial service, investment bank, percentage point, price rise, consumer product, prescription drug, high profile, profit margin, final round, people live, police officer, meet need, plan build, chip maker, middle class, home builder, recent survey, trucking company, city official, grow market, emerge market, executive director, investment officer, federal judge, make sense, price earning, computer maker, high court, work hard, people close, broad market, executive chairman, family member, company fund, double digit, human resource, prime minister, comic book, company serve, personal computer, central bank, attorney general, product development, employ people, expect generate, company offer, company business, global head, local government, golf course, health plan, capital spending, look forward, energy price, grow number, hand hold, tell investor, transaction expect, accord court, computer chip, business practice, merger acquisition, support service, shopping mall, exercise option, market expect, service group, asset manager, management business, foreign company, good time, people work, include work, institutional investor, accord report, maker sell, private sector	share accord, company share, clinical trial, revenue fall, share rise, share jump, company stock, share company, expect earning, bankruptcy protection, stock fall, price share, sale fall, trading stock, share share, company plan, patient receive, stock rise, sale company, share cash, report earning, regulatory filing, drug call, file bankruptcy, post loss, analyst expectation, bankruptcy court, gain share, familiar matter, drug company, person familiar, share analyst, income share, earning share, closing price, share revenue, company sell, composite trading, revenue company, short seller, company deal, estimate analyst, income rise, compare loss, drug market, chief executive, fall company, biotechnology company, board member, develop drug, company board, report revenue, company statement, market close, biotech company, drug development, estimate share, restructuring plan, cash stock, deal company, result include, send share, share outstanding, operating profit, earning report, increase revenue, expect report, share plunge, agree acquire, intellectual property, initial public, share close, company chief, insider trading, compare share, fall sharply, treat patient, product sale, report profit, drug maker, expect revenue, increase share, company profit, share period, public company, cancer drug, financial result, save money, cancer treatment, mortgage lender, slot machine, company begin, analyst expect, expect result, sale sale, lose value, market capitalization, suit file, expect company, share base

Table A9
Dictionaries for Option Return Determinants

Table A9 lists important word features for various option return determinants from the support vector regression (SVR) model. Option return determinants are the same as those mentioned in Table 8. Important word features are defined as the top 1000 terms with the largest magnitudes (i.e., the absolute value of the coefficients) from the SVR model. For abbreviation, we list the first 100 terms with the highest frequency of positive or negative occurrences. The sample period is from January 1996 to November 2022.

	Positive	**Negative**
ΔIV	movie, rise, deliver, sale, revenue, strong, computer, growth, dollar, video, forecast, maker, single, business, exclude, analyst, average, lift, price, bump, income, good, litigation, strengthen, online, subscription, publisher, comparable, gross, grow, presence, high, complaint, earning, rival, report, drive, weak, profit, house, digital, increase, margin, downtown, card, housing, kitchen, feed, star, share, wireless, version, upgrade, traveler, shopping, modest, article, expectation, loss, result, reflect, game, retailer, double, chain, period, fiber, post, coffee, school, hire, stake, store, expensive, nearby, vacation, service, boost, product, packaging, appeal, complete, interesting, antitrust, seat, cosmetic, designer, course, cite, surround, software, plaintiff, capture, metal, seasonal, solid, transport, property, memory, host	schedule, payment, exposure, immediately, evaluate, sheet, spokesman, regulation, bond, governor, call, serve, give, level, option, trader, cancel, buyout, sell, detail, pharmacy, intend, review, severe, selloff, warning, shareholder, fund, study, seek, great, announce, public, hold, award, cell, obligation, administration, necessary, accord, asset, uncertainty, crisis, generic, assume, damage, government, salary, confirm, recognize, recession, bank, lose, head, start, respond, surprise, take, chemical, contract, flow, therapy, dealer, receive, consult, meeting, commitment, effort, seriously, nation, likely, director, manage, support, buyback, program, appropriate, aware, equity, grant, volatility, drug, representative, exist, health, marketer, current, block, repair, troop, work, hearing, agency, manager, medical, electric, oversee, derivative, hedge, debt
IVOL	share, company, stock, loss, revenue, trading, online, tumble, news, fall, analyst, plunge, cell, close, short, announce, music, closing, genetic, rental, trial, passenger, report, sharply, flier, clinical, cash, listing, surge, founder, inventory, soar, flight, digital, base, biotech, chief, airport, drop, forecast, booking, disappointing, airline, destination, traveler, announcement, retailer, bankruptcy, seller, jump, attendant, biotechnology, treatment, patient, user, credit, audio, plummet, casino, gambling, traffic, subscriber, capitalist, enrollment, internet, expect, result, flash, gene, amortization, clothing, selling, royalty, download, store, drug, apparel, mall, site, fashion, expectation, widen, video, filing, carrier, mortar, subscription, downgrade, investor, halt, slump, mail, rent, price, fare, outlook, disease, tech, killing, warrant	global, head, division, unit, currency, vice, retirement, packaging, boost, spokesman, foreign, commercial, diaper, railroad, investment, meeting, beverage, mutual, defense, marketer, manager, insurance, rate, asset, client, return, industrial, career, overall, join, conglomerate, multinational, environmental, corporation, train, policy, cigarette, risk, dollar, bank, bond, banking, snack, oversee, business, retire, cereal, dividend, consumer, inflation, director, household, tobacco, fine, detergent, flavor, graduate, insurer, freight, paint, corporate, water, aerospace, appeal, spokeswoman, strong, program, settlement, account, institution, total, building, plaintiff, plant, president, drink, fund, growth, giant, move, label, food, chocolate, sponsor, chemical, article, banker, management, branch, budget, sugar, rule, include, paper, nuclear, tournament, approach, portfolio, supermarket, innovation

HV-IV	share, revenue, report, user, rise, beat, deal, analyst, stock, gross, maker, closing, trading, equipment, expectation, technology, acquisition, premium, compete, compare, price, surge, forecast, cash, base, computer, gain, agree, software, regulatory, offer, high, value, opportunity, customer, offering, help, function, chip, period, enable, rebound, slightly, device, common, storage, demand, partner, service, well, jump, combined, company, personal, data, firm, listing, profit, merger, semiconductor, rival, system, page, approval, strong, prior, datum, site, infrastructure, earning, release, networking, shareholder, chart, growth, science, margin, announce, transaction, prepared, capital, sale, positive, head, predict, partly, agreement, expand, miss, downturn, climb, integrate, post, multiple, income, active, boom, result, player, list	downgrade, auto, meat, affect, tobacco, material, retire, director, president, cosmetic, retailer, beer, tight, explore, promotion, game, holiday, pressure, labor, department, station, loan, blood, cigarette, chairman, exposure, estate, safe, supermarket, spokesman, fire, matter, crisis, food, spokeswoman, hedge, brand, production, measure, flight, contractor, cereal, sleep, friend, paint, patent, financial, corporate, shelf, borrower, newspaper, buyback, possible, publish, marketer, club, warehouse, notice, medication, face, back, session, stroke, turmoil, sweater, chain, owner, generic, beverage, lender, default, poultry, professor, governor, smoking, affordable, abandon, effect, dairy, replace, strike, retail, schedule, staff, real, rail, builder, satellite, campaign, specialty, present, risk, think, brewer, warning, label, favorable, pork, familiar, letter
ILLIQ	loss, firm, debt, contemporary, news, short, dear, swim, vehicle, fiction, liquidity, publicly, playwright, resign, voter, choreographer, peace, chicken, assessment, prisoner, technology, gene, hope, collapse, closure, rental, airport, inmate, consolidation, alternative, inquiry, murder, dancer, writer, firearm, tire, fare, tumble, outflow, musician, plummet, online, bureau, piano, photography, adore, boat, staging, music, rent, estate, diplomatic, public, make, faculty, athletic, real, escape, composer, dairy, poultry, print, plunge, adviser, unchanged, survivor, lose, passenger, independent, base, terrorist, pool, furniture, terrace, historian, referendum, publish, listing, credit, steel, racing, recording, indict, photograph, commander, collection, guitar, buyer, choreography, stone, exhibition, chief, widen, royal, found, vocal, artist, unsolicited, fireplace, drop	giant, unit, insurer, chip, earning, spokesman, innovation, pharmacy, currency, global, multinational, diaper, license, pipeline, corporation, foreign, project, conglomerate, buyback, computing, world, bank, acquisition, franchise, drug, pharmaceutical, debit, spokeswoman, article, equipment, railroad, cable, tech, regulator, growth, join, processing, multiple, refining, toothpaste, employee, card, analyst, carry, warning, official, engine, utility, engineer, plant, packaging, blockbuster, arthritis, warehouse, repurchase, globally, benefit, rival, program, income, broad, exclude, vice, field, detergent, promote, cereal, storage, titan, strong, grocery, soap, networking, sponsor, initiative, throw, increase, prescription, rise, aerospace, hotel, paint, cigarette, improvement, overall, telecommunication, sugar, package, gasoline, implant, innovate, retire, insurance, cancer, hardware, banking, chocolate, successor, energy, index

MFIS	download, audio, delete, miner, browser, subscription, desktop, internet, graphic, user, click, content, gaming, virtual, hacker, server, computer, genetic, streaming, software, video, copy, gambling, online, digital, wireless, programming, keyboard, taxi, sequence, mortar, scan, search, lung, subscriber, memory, censor, interface, chat, recording, publisher, console, computing, copyright, spam, disk, brick, gamer, music, clinical, flier, cellular, videogame, vacation, bookstore, reviewer, tourist, password, blog, programmer, geek, wrist, voucher, defenseman, capitalist, pizza, creator, gadget, nonstop, screen, antitrust, stream, royalty, useful, library, broadband, browse, random, scroll, cell, conductor, visual, hack, rumor, commerce, headphone, mouse, delight, portable, listener, worm, classmate, exploit, messaging, carrier, steel, laptop, inspector, virus, viewing	credit, firm, state, portfolio, bond, investment, client, insurance, manage, lender, banking, asset, bank, rate, fund, move, risk, economist, debt, yield, create, plan, accord, director, people, increase, insurer, article, dividend, lending, level, economic, money, member, inflation, economy, equity, unit, group, health, represent, market, agency, issue, employee, private, tell, interest, loan, high, operation, institution, industrial, want, hold, sell, number, recently, percentage, capital, mutual, need, index, benefit, payment, begin, policy, come, consider, area, financial, crisis, deposit, case, effort, account, overall, office, federal, finance, good, default, emerge, balance, life, join, rise, exchange, official, face, broker, income, saving, include, commercial, senior, total, nation, lead, utility
MFIK	food, tobacco, group, retire, cigarette, acquisition, utility, unit, director, bank, come, include, banking, pension, rate, deal, agency, asset, snack, division, fund, increase, degree, packaging, yield, portfolio, takeover, institution, cereal, manager, base, beverage, branch, business, acquire, lending, value, time, mortgage, regulatory, graduate, income, borrower, dividend, line, state, offer, change, remain, management, insurance, smoker, hold, drink, retirement, head, debt, saving, country, bond, owner, transaction, interest, power, regulator, electricity, inflation, consumer, investment, private, score, agree, water, family, loan, smoking, defense, high, property, area, premium, long, team, system, know, president, home, role, life, shareholder, father, brand, risk, meat, accord, plant, chairman, work, regional, product	internet, software, download, chip, click, tech, computer, semiconductor, server, website, screen, clothing, browser, search, video, online, desktop, button, subscription, inventory, user, surge, cell, plunge, browse, virtual, drilling, networking, digital, demand, technology, passenger, audio, booking, flier, outfit, shirt, storage, site, airline, boom, graphic, delete, cabin, miner, attendant, profitability, mining, shopper, traffic, capitalist, videogame, mobile, interface, offering, code, destination, shopping, computing, stock, commerce, hardware, crude, carrier, console, camera, steel, biotech, tool, wireless, production, entrepreneur, memory, mouse, loss, plug, gaming, mortar, flight, mall, developer, function, equipment, disk, traveler, portal, data, visa, founder, sharing, engine, quickly, keyboard, display, genetic, patent, network, speaker, capacity, builder

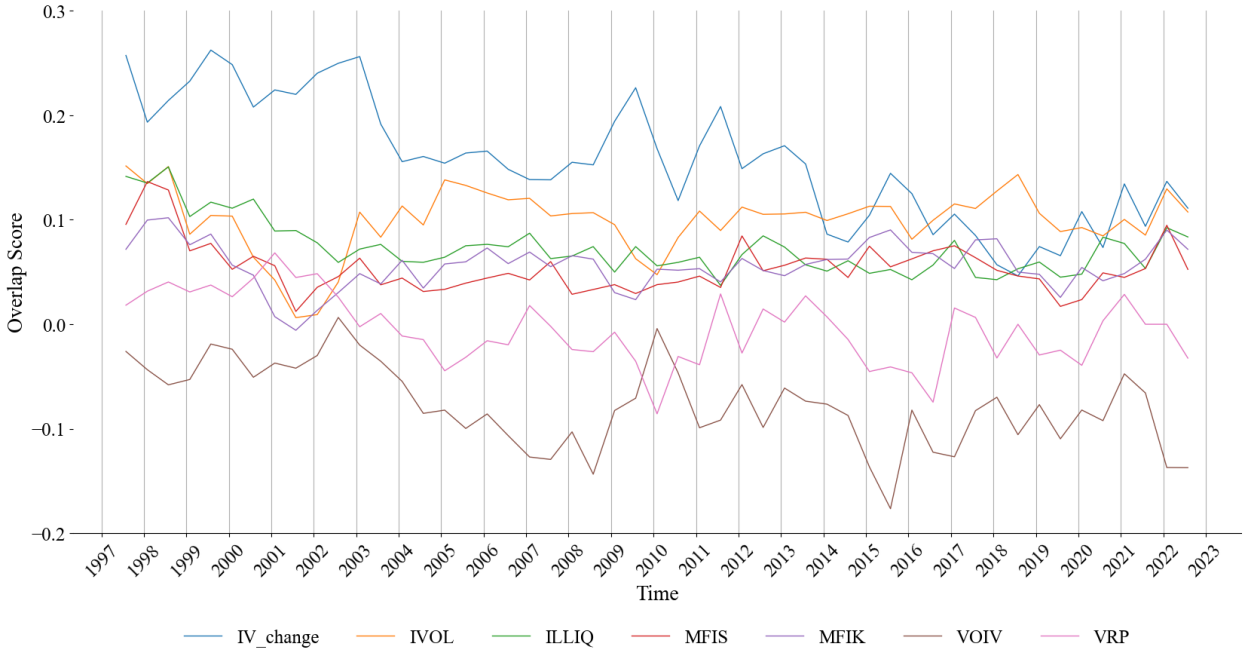
VOIV	share, forecast, trading, revenue, plunge, earning, credit, debt, stock, announce, tumble, fall, utility, volatility, loss, hurt, income, jump, deal, meat, disappointing, slide, weakness, affect, close, outlook, song, plummet, sink, takeover, transaction, combine, selloff, loan, value, expense, default, reject, conference, telecom, restructuring, result, short, fraud, face, cash, shareholder, climate, interview, miss, projection, fear, company, surge, acquisition, downgrade, slip, investigator, financing, uncertainty, imply, news, buyout, anticipate, volatile, cloud, lower, expectation, speculation, bank, speed, provider, exclude, range, charge, overseas, letter, representative, spending, trouble, software, weak, wake, turmoil, merger, caution, bankruptcy, hostile, reaction, interim, electricity, freight, membership, book, premium, food, guidance, profit, review, spark	drill, production, builder, seat, field, want, supply, foreign, contract, land, barrel, export, boost, exploration, refining, crude, gift, aircraft, drilling, taxpayer, oversee, military, packaging, bonus, missile, issuance, need, defense, mining, aluminum, experience, improve, refiner, price, national, steel, physical, partnership, manufacturer, produce, building, unit, airline, brokerage, benefit, offering, advantage, gasoline, truck, joint, gain, partly, policy, inch, site, miner, recover, refinery, interest, exemption, presentation, issue, vehicle, rocket, tool, terminal, aerospace, plant, dominant, extract, online, administration, lift, traffic, fuel, restrict, participate, increase, wear, chemical, fighter, look, demand, tanker, employer, plane, broker, individual, president, producer, potential, soldier, rally, suit, league, goal, passenger, diesel, award, planning

Table A10
An Example of the News Article

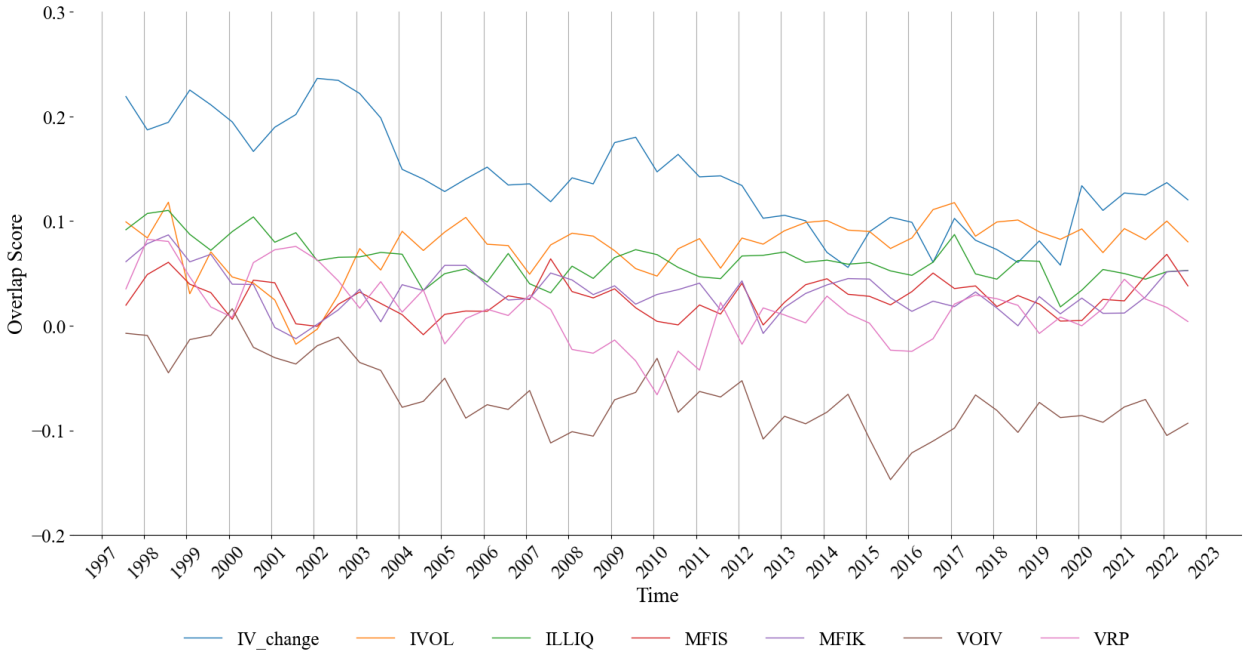
Table A10 shows a news article example included in our sample. As an example, we highlight words that positively (negatively) forecast delta-hedged call option returns with red (blue).

Tesla Inc. stock has soared, pushing the company's market value over $100 billion. Its bonds, however, are a whole lot calmer. Tesla's bond maturing in 2025 traded recently at 99.50 cents on the dollar, little changed from 97 cents since the start of the year. Shares have risen 39% so far in 2020 and surged more in off-hour trading after the company reported results that exceeded Wall Street analysts' expectations Behind the relative quiet in bonds: Investors there tend to care less about the company's long-term growth prospects than what happens to its roughly 10 billion of debt if it defaults or declares bankruptcy. That leaves them digging into cash flows to understand how even a crippled Tesla could raise cash, keep the lights on and repay creditors. Most bond investors and analysts believe they would receive full repayment on the electric-car maker's debt even in a bankruptcy. Even if Tesla were to go bust, a last-resort buyer would likely buy the company for at least $10 billion, several investors said. The company has numerous assets that would appeal to other auto makers, they said, including a factory in Nevada, intellectual property, the Tesla brand and the company's lead in battery technology.

Figure A1. Time Series of the Overlap Score for each Option Return Determinant



(a) Call



(b) Put

Figure A1: This figure shows the time series of the overlap score for each option return determinant.