Forecasting Option Returns with News*

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

February 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 their delta-hedged 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 understand the underlying sources of our textual predictors. We find that the predictive power of the textual predictors stems from a composite effect, with future implied volatility changes being the most decisive, alongside significant contributions of various other option return determinants. Our study highlights the importance of analyzing text data using machine learning approaches to forecast option returns.

Keywords: textual analysis, news media, return predictability, machine learning, delta-hedged options

JEL classification: G12, G13, G14, G17

*We thank Turan Bali, Svetlana Bryzgalova, Hector Chan, Amit Goyal, Evan Jo, Mete Kilic, Hugues Langlois, Asaf Manela, Paul Tetlock, Sang-Ook Shin, Yanchu Liu, Laurence van Lent as well as seminar participants at Nankai University, CFE-CMStatistics (2021), CICF (2022), AsianFA (2022), CIRF (2022), SFS Cavalcade Asia-Pacific (2022), FERM (2023), APAD (2023), and China Derivatives youth Forum (2023). The work described in this paper is supported by the 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), and Frankel, Jennings, and Lee (2022), highlights that machine-learning methods can yield more potent and reliable asset pricing implications compared to dictionary-based approaches.

Despite the rich application of text data on the equity market, there is limited knowledge about the implications of textual analysis in the option market, particularly equity options.[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, 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, Stroebel, and Karolyi (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] Manela and Moreira (2017) construct a text-based measure of uncertainty using support vector regression to construct a news-based VIX before 1990, although they only focus on index options.

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 0.49% (0.33%) 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, and neural networks, and to different constructions of word features.

Furthermore, we examine potential mechanisms by which news media forecast delta-hedged option returns. We propose a novel method to explore what information from news media drives such return predictability. In particular, we first create a dictionary containing important words that help to forecast option returns based on the feature importance through our machine-learning model. Subsequently, we implement a similar process and project different option return determinants, such as implied volatility change, idiosyncratic volatility, volatility deviation, illiquidity measure, 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, implied skewness, and implied kurtosis. 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. Additionally, 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.

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 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.

[4]See also Ramachandran and Tayal (2021), Choy and Wei (2022), Jeon, Kan, and Li (2019) for recent developments in this literature

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 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 these 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. A textual fuzzy match, such as Jaro-Winkler distance or Levenshtein distance, is applied to define how similar a specific string is to the target string. We then assign each 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 be identified 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. Prior to 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. We also collect variables on

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

underlying stocks, such as stock return, stock price, trading volume, and shares outstanding, from the CRSP database. Our sample period covers 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,[6] 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). Finally, we retain only those options with a moneyness ranging from 0.8 to 1.2.[7] All filters are strictly based on information prior to the portfolio formation date, so that 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 1,010,845 article-month observations and 828,878 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 88,630 option-month observations for both call and put options on individual stocks over a 323-month sample period from January 1996 to November 2022. On average, we have 274 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 3.64% of the total number of firms in the market, the total market cap of these firms represents 33.39% of the total market. In the universe of optionable stocks, our sample comprises 9.38% of the total number and 35.93% of the total market capitalization of optionable stocks, on average. As shown in Table A1 in the Appendix, firms in our sample rank the 87^{th} percentile on average in the CRSP stock universe with an average firm size of 32.63 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 an approximately representative of the whole market.

[6] \$0.10 for options trading above \$3 and \$0.05 otherwise

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

2.2 Variable Constructions

The main outcome variable of our study is returns to delta-hedged options with daily rebalancing, which is the dollar gains of daily-rebalanced delta-hedged long option positions scaled by the initial costs.[8] The delta-hedged call option gain is defined as 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. 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 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 puts. Hence, our delta-hedged option returns are defined as:

$$r_{i,t}^{\text{call}} = \frac{\Pi_{t,t+\tau}}{\Delta_{c,t} S_t - C_t} \tag{2}$$

$$r_{i,t}^{\text{put}} = \frac{\Pi_{t,t+\tau}}{P_t - \Delta_{p,t} S_t} \tag{3}$$

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 frequency. Since we train our model on a rolling basis, we repeat the same process independently for each training process instead of applying this procedure to the entire corpus. This ensures that our textual predictors are devoid of future information, relying solely on data available at

[8] Tian and Wu (2021) show that the daily-rebalanced delta-hedging strategy can remove as high 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).

the training stage. Appendix Table A2 demonstrates that the top 10,000 unigrams already represent over 95% of the unigrams 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 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

$$w_{i,t}^{j,tfidf} = \begin{cases} 1 + \log(tf_{i,t}^{j}) w_t^{j,idf}, & if \;\; tf_{i,t}^{j} > 0 \\ 0 & otherwise \end{cases} . \tag{4}$$

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 proposed by Manela and Moreira (2017), we consider the following linear regression problem in cross-section at the end of each month:

$$r_{i,t} = \alpha_t + \beta_t' x_{i,t-1} + \epsilon_{i,t}, \qquad i = 1, 2, \ldots, N_t, \tag{5}$$

where $r_{i,t}$ is delta-hedged option returns for firm i over the month $[t-1, t]$, and $x_{i,t-1} = [x_{i,t-1}^1, \ldots, x_{i,t-1}^K]'$ 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:

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

$$subject\ to \begin{cases} r_{i,t} - \beta_t' x_{i,t-1} - \alpha_t \leq \epsilon + \xi_{i,t} \\ \beta_t' x_{i,t-1} + \alpha_t - r_{i,t} \leq \epsilon + \xi_{i,t}^* \quad , i = 1, 2, \ldots, N_t. \\ \xi_{i,t}, \xi_{i,t}^* \quad \geq 0 \end{cases} \tag{6}$$

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 larger than ϵ are tolerated. This corresponds to dealing with the so-calledr ϵ- insensitive loss function $|\xi|_\epsilon$ described by:

$$|\xi|_\epsilon := \begin{cases} 0, & if \ |\xi| \leq \epsilon \\ |\xi| - \epsilon, & otherwise \end{cases} . \tag{7}$$

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 spanning one year, followed by a validation sample of six months and a test sample of six months. The training set encompasses 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 SVR in the test sample is considered as the textual information extracted from news media pertaining to 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:

$$TP_{i,t} \equiv \hat{r}_{i,t+1} = \hat{\alpha}_t + \hat{\beta}_t' x_{i,t}, \qquad i = 1, 2, \ldots, N_t. \tag{8}$$

where $\hat{\alpha}_t$ and $\hat{\beta}_t$ are fitted parameters in Equation (2) based on SVR 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]

Table 1 Panel B reports the time-series average of the cross-sectional correlations between ML textual predictors and many existing option return determinants. Although ML textual predictors have relatively high correlations with each other (0.66), 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 the high-dimensional data of 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, 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 Call_SVR and Put_SVR, 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 subsequent month.

[Insert Table 2]

Table 2 demonstrates that textual information predicts delta-hedged equity option returns. The monthly long-short strategy based on textual information generates significant return spreads economically and statistically. For example, the average monthly return spread between the top and the bottom quintiles sorted by the textual predictors using SVR is 0.49% (0.33%) for call (put) options. This return spread is substantial, accounting for 119.51% (61.11%) of the absolute value of the median of the daily-rebalanced option return, which is 0.41% (0.54%) for call (put) options. We also adjust the portfolio return spreads

using two factor models. The first factor model is a senven-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 factor model is the option two factor model from Zhan et al. (2022), including an idiosyncratic volatility factor and an illiquidity factor. Our results are robust to the risk adjustments, and the adjusted alphas match the original return spreads closely, indicating that common risk factors do not drive our results.

As shown in Table 2, textual information predicts delta-hedged option returns for at least one month, while it predicts stock returns only for a few days. 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 for the stock market. The persistence of option return predictability could be driven by the corresponding persistency of volatility-related information. In Section 4, we confirm our hypothesis and show that there is considerable overlap between the textual information pertinent to delta-hedged option returns and that associated with changes in implied volatility.

3.1.2 Double Portfolio Sorting

To examine whether the effects of our ML 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 predictor. 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 illiquidity measure (ILLIQ), (4) Uncertainty about the implied volatility (VOIV), calculated as the standard deviation of the daily percentage change of option implied volatility during the month; (5) Model-free implied skewness (MFIS), calculated as in Bakshi and Kapadia (2003);(6) Model-free implied kurtosis (MFIK), calculated as in Bakshi and Kapadia (2003);

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 five portfolios based on our ML textual predictors (i.e., Call_SVR or Put_SVR). 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 ML textual predictors. The return spreads of call options range from 0.42% to 0.48% per month, and those of put options range from 0.25% to 0.34% 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 the 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 7-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 ML 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 ML textual predicator:

$$r_{i,t} = \alpha_t + \beta_t TP_{i,t-1} + \sum_{j=1}^{M} \gamma_t^j X_{i,t-1}^j + \epsilon_{i,t}, \quad i = 1, \ldots, N_t, \tag{9}$$

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 $X_{i,t-1}^j$ 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 one standard deviation.

We run the cross-sectional regression of Equation (9) each month. After obtaining the time series of the coefficients (e.g., β_t) for the independent variables, we conduct 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 ML 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 Call_SVR (Put_SVR) is 0.17 (0.11) with t-statistics of 5.38 (5.12). In a multivariate regression, the coefficients on the ML textual predictors remain economically and statistically significant, with a coefficient on Call_SVR (Put_SVR) of 0.11 (0.06) and a t-statistic of 3.85 (2.86). 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 stock volatility deviation has a positive effect on forecasting the cross-section of equity option returns. In Table A3 in the appendix, we incorporate additional controls for dictionary-based measures for sentiment and uncertainty derived from the Loughran-McDonald dictionary. Table A3 shows that these measures exhibit marginal significance when entering the regression, and the inclusion of dictionary-based measures can hardly affect the coefficients of our ML textual predictors.

3.2 Robustness Checks

3.2.1 Alternative Machine Learning Approaches

So far, we have demonstrated the usefulness of using 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 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 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 applies 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.[9] We conduct robustness checks using different parameters of C and ϵ to train the SVR model. We show in an

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

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 to different input variables for the machine learning model. For instance, in constructing the tf-idf matrix, we experiment with various number 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 training period 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 rolling window duration.

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 for news media for predicting option returns.

A model choice close to SVR is 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 in the following way:

$$\alpha_t, \beta_t = \underset{\alpha_t \in R, \beta_t \in R^K}{\arg\min} \Big\{ \frac{1}{N_t} \sum_{i=1}^{N_t} \Big(r_{i,t} - \alpha_t - \sum_{k=1}^{K} \beta_t^k x_{i,t-1}^k \Big)^2 + \lambda \sum_{k=1}^{K} |\beta_t^k| + (1-\lambda) \sum_{k=1}^{K} (\beta_t^k)^2 \Big\}, \quad (10)$$

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 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:

$$\hat{r}_{i,t+1} = \hat{\alpha}_t + \sum_{k=1}^{K} \hat{\beta}_t^k x_{i,t}^k, \qquad i = 1, \ldots, N_t. \tag{11}$$

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. While 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 important for predicting option returns using textual information because words are heavily 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.

The 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 can 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:

$$\hat{r}^{RF}(x) = B^{-1} \sum_{b=1}^{B} \bar{r}(T_b^*(x)), \tag{12}$$

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 networks, 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 the layers in between these 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:

$$h_i^{(1)} = \phi^{(1)}\Big(\sum_j w_{ij}^{(1)} x_j + b_i^{(1)}\Big), \tag{13}$$

$$h_i^{(2)} = \phi^{(2)}\Big(\sum_j w_{ij}^{(2)} h_j^{(1)} + b_i^{(2)}\Big), \tag{14}$$

$$r_i = \phi^{(3)}\Big(\sum_j w_{ij}^{(3)} h_j^{(2)} + b_i^{(3)}\Big), \tag{15}$$

The nonlinearity and interaction among words can be captured by the nonlinear activation functions and 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.

[Insert Table 5]

To save space, Table 5 presents a single portfolio sorting 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 0.72 with SVR textual predictors, while the neural network and random forest textual predictors have correlation coefficients of 0.60 and 0.52 with the SVR textual predictors, respectively.

3.2.2 Alternative Constructions of Word Features

In our main analysis, we train the machine learning models using unigrams word counts (adjusted by document frequency), because of its simplicity and effectiveness. However, it has two main 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 unigrams 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 bigram, trigram, or n-gram. 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 (3) and construct the corresponding textual predictor based on Equation (8). Given the fact that the total number of features increases exponentially when switching from unigrams to n-grams ($n > 1$), we include more features into 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 test 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 21.34% of the total bigram features, and the top 80,000 trigram tokens cover only 5.41% of the total trigram features, as shown in Table A2. Therefore, the return spreads in Table 6 are 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 Time Series of Return Spreads and Subperiod Analysis

Avramov, Cheng, and Metzker (2023) documents that trading strategies utilizing machine learning yield higher profits in periods of 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 machine-learning textual predictors across different market conditions. 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) 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 machine-learning (ML) textual predictors across different market conditions. Table 7 shows that the return spreads generated by ML textual predictors are robust in different market conditions. Furthermore, aligning with the 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. In addition, the return spreads are also significantly higher during recession periods.

[Insert Table 7]

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.

In this section, we propose a novel method to offer insights into 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 2000 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:

$$\text{Positive Score} = \frac{P}{P+N} \tag{16}$$

$$\text{Negative Score} = \frac{N}{P+N} \tag{17}$$

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, given 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 in the Appendix displays the top 100 words 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 include 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 48.5% (47.7%). Conversely, only 6.5% of the positive words in the call option dictionary overlap with negative words in the put option

dictionary, and 4.7% 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.

4.1.2 Topic Analysis of Option Return Dictionaries

While important word features provide some insights, they can be too detailed and not easy to understand at first. This section aims to sort these word features into 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 weights to each word based on its relevance to each topic. We apply these weights to our option return dictionary. Specifically, for each word in our option return dictionary, we obtain its weights for each of the 180 topics. We then aggregate these weights for each dictionary and rank the topics by their total weights, from highest to lowest.[10]

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 *mining*, *economic growth*, and *financial crisis* show a positive correlation with call option returns, whereas *steel*, *job cuts*, and *small business* align positively with put options. Conversely, firm-specific or event-driven topics tend to negatively impact option returns. Notably, among the top 10 topics, *M&A*, *Takeovers*, *People familiar*, *IPOs*, *Share payouts*, *Exchanges/composites*, *Earnings losses*, and *Negotiations* are negatively associated with future equity option returns. Additionally, there is more overlap in topics negatively related to option returns: four topics overlap between the positive dictionaries for call and put returns, while eight topics overlap in the negative dictionaries.

[Insert Figure 2]

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

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 ML textual predictors. Table 4shows that the inclusion of various ex-ante determinants of expected option returns leads to a substantial decrease in the coefficients of our ML textual predictors (35.29% for call options and 45.46% for put options), indicating some overlap in the information content of the textual predictors and these option return determinants. We hypothesize that news 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 illustrating 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 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 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 A8 in the Appendix 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:

$$\text{Overlap Score} = \frac{C - W}{C + N + W} \tag{18}$$

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. 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 42.5%. Similarly, the overlap score between MFIS and MFIK stands at 37.65%.

[Insert Figure 4]

[Insert Table 8]

We calculate the overlap score between option return dictionaries and six option return determinant variables, and present the results in Table 8. Among various option return determinants, the highest average overlap score is observed for the change of implied volatility, which has an average overlap score of 19.85% (18.7%) for the call (put) option dictionary.[11] This suggests that the information set related to change of implied volatility is a key source of

[11] We confirm this finding in a time series setting. Specifically, in each training iteration, we compute the

textual information for option return predictability. Other option return determinants that have relatively high overlap scores include idiosyncratic volatility (17.5% for call and 16.6% for put), model-free implied skewness (12.65% for call and 8.3% for put), and model-free implied kurtosis (11% for call and 9.1% 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 exhibit meaningful overlap scores, indicating their contribution to the predictability of the textual information.

Table 8 shows the advantage of the ML 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 provide the most important contribution when forming the SVR 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:

$$\mathrm{TP}_{i,t} = \hat{\alpha}_t + \sum_{k=1}^{K} \hat{\beta}_t^k x_{i,t}^k, \quad i = 1, \ldots, N_t \tag{19}$$

overlap score for every option return determinant and draw the pattern of these overlap scores. Figure A1 in the appendix shows the time series pattern of overlap scores, with the overlap score of the change of implied volatility consistently ranking the highest most of the time.

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 9]

Table 9 shows that change of implied volatility has the highest coefficient and adjusted 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 an untabulated table, we confirm that change 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.[12] 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 be quantified through a predefined dictionary. While information in news contents related to change of implied volatility is the major source of the predictability of our textual predictors, it is noteworthy that the fitted values of change of implied volatility do not exhibit comparable predictability for delta-hedged option returns. In an untabulated table, we verify that the fitted values of 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.

5 Conclusion

In this paper, we study whether and how textual information from news media can be used to enhance option return predictability. First, we find that news media contains substantial information for future delta-hedged option returns. The results are robust after

[12]The sign of the fitted value of the model-free implied kurtosis flipped because it captures similar information with the model-free implied skewness, and the sign of HV-IV is more negative in the horse-racing specification, reassuring its little contribution to the formation of our textual predictors

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 constructing word features. Our results demonstrate that 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 implied skewness, idiosyncratic volatility, and implied kurtosis. 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 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.

Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, The cross-section of volatility and expected returns, *The 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, *The 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, *The Review of Financial Studies* 36, 3548–3602.

Boulatov, Alex, Assaf Eisdorfer, Amit Goyal, and Alexei Zhdanov, 2022, Limited attention and option prices, *SSRN Electronic Journal* .

Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu, 2021, The structure of economic news, *National Bureau of Economic Research Working Paper Series* 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?, *The Quarterly Journal of Finance* 13.

Chinco, Alex, Adam D. Clark-Joseph, and M. A. O. Ye, 2019, Sparse signals in the Cross-Section of returns, *The Journal of Finance* 74, 449–492.

Choy, Siu Kai, and Jason Wei, 2022, 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, *The 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, X. I., Y. A. N. Li, David E. Rapach, and Guofu Zhou, 2021, Anomalies and the expected market return, *The Journal of Finance* 77, 639–681.

Engle, Robert F., Stefano Giglio, Bryan Kelly, Heebum Lee, Johannes Stroebel, and Andrew Karolyi, 2020, Hedging climate change news, *The 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.

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

Gu, S. H., B. Kelly, and D. C. Xiu, 2020, Empirical asset pricing via machine learning, *Review of Financial Studies* 33, 2223–2273.

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.

Jeon, Yoontae, Raymond Kan, and Gang Li, 2019, Stock return autocorrelations and the cross section of option returns, *SSRN Electronic Journal* .

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, T. I. M., and Bill McDonald, 2011, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *The 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.

Newey, Whitney K., and Kenneth D. West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica : journal of the Econometric Society* 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.

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, *The 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, *The Journal of Finance* 63, 1437–1467.

Tian, Meng, and Liuren Wu, 2021, Limits of arbitrage and primary risk taking in derivative securities, *SSRN Electronic Journal* .

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

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

Table 1
Summary Statistics

This table 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. *Call_SVR* (*Put_SVR*) is the textual predictor extracted from news media for call 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 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 (%)	-0.07	5.56	-3.88	-1.92	-0.41	1.16	3.61
Put Option Return (%)	-0.38	4.33	-3.96	-2.02	-0.54	0.94	3.12
Call_SVR (%)	-0.42	0.26	-0.76	-0.59	-0.41	-0.24	-0.10
Put_SVR (%)	-0.55	0.27	-0.90	-0.72	-0.54	-0.37	-0.22
IVOL (%)	1.93	1.35	0.86	1.12	1.57	2.32	3.35
HV-IV (%)	1.60	10.14	-7.36	-2.76	1.15	5.44	11.21
ILLIQ	-8.41	1.57	-10.32	-9.49	-8.55	-7.38	-6.28
MFIS	-0.49	0.44	-0.98	-0.70	-0.47	-0.26	-0.04
MFIK	4.40	1.36	3.26	3.52	3.99	4.86	6.11
VOIV (%)	5.91	4.22	3.17	3.91	5.00	6.59	8.92

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

	Call_SVR	**Put_SVR**	**IVOL**	**HV-IV**	**ILLIQ**	**MFIS**	**MFIK**	**VOIV**
Call_SVR	1	0.66	-0.09	0.01	-0.09	-0.05	0.01	-0.03
Put_SVR		1	-0.09	0.01	-0.08	-0.04	0.00	-0.04
IVOL			1	0.10	0.35	0.19	-0.20	0.21
HV-IV				1	-0.05	-0.06	0.08	-0.03
ILLIQ					1	0.16	0.06	0.12
MFIS						1	-0.36	-0.04
MFIK							1	0.19
VOIV								1

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

This table reports the average monthly excess returns to the delta-hedged option portfolios sorted by *Call_SVR* (*Put_SVR*). At the end of each month, we rank all underlying stocks into quintiles by their *Call_SVR* (*Put_SVR*). Detailed descriptions of *Call_SVR* (*Put_SVR*) are provided in Section 2.2. The portfolio is held for one month. 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 a seven-factor model and report the corresponding alphas. The 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 option two factor model includes an idiosyncratic volatility factor and an Illiquidity factor as described in Zhan et al. (2022). The realization of 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 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	-0.31	-0.14	-0.02	0.05	0.18	0.49***
		(-2.56)	(-1.07)	(-0.13)	(0.44)	(1.30)	(5.36)
	7-Factor α	-0.22	-0.03	0.04	0.14	0.23	0.45***
		(-1.41)	(-0.21)	(0.24)	(0.88)	(1.27)	(4.04)
	Option 2-Factor α	-0.35	-0.19	-0.08	-0.01	0.11	0.46***
		(-3.39)	(-1.74)	(-0.67)	(-0.09)	(0.86)	(4.60)
		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 – 1)**
Put	Average Return	-0.56	-0.39	-0.35	-0.32	-0.23	0.33***
		(-6.34)	(-4.10)	(-3.70)	(-3.60)	(-2.35)	(4.92)
	7-Factor α	-0.57	-0.37	-0.34	-0.32	-0.22	0.35***
		(-5.57)	(-3.67)	(-3.28)	(-3.51)	(-2.14)	(4.64)
	Option 2-Factor alpha	-0.64	-0.46	-0.40	-0.39	-0.31	0.33***
		(-6.76)	(-4.23)	(-4.09)	(-4.15)	(-3.15)	(4.94)

Table 3
Dependent Double Protfolio Sorting

In this table, we investigate whether several control variables can individually explain the effect of ML 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 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 ML-based 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-West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Call Options	Put Options
Baseline	0.49^{***}	0.33^{***}
	(5.36)	(4.92)
IVOL	0.48^{***}	0.29^{***}
	(5.91)	(4.65)
HV-IV	0.48^{***}	0.29^{***}
	(5.08)	(4.26)
ILLIQ	0.42^{***}	0.25^{***}
	(5.26)	(4.54)
MFIS	0.45^{***}	0.31^{***}
	(5.39)	(4.51)
MFIK	0.45^{***}	0.32^{***}
	(5.09)	(5.06)
VOIV	0.46^{***}	0.34^{***}
	(5.11)	(4.94)

Table 4
Fama-MacBeth Regressions

This table reports the Fama-Macbeth regression results of the delta-hedged equity option returns on ML 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-West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Call		**Put**	
	(1)	**(2)**	**(3)**	**(4)**
SVR	0.17***	0.11***	0.11***	0.06***
	(5.38)	(3.85)	(5.12)	(2.86)
IVOL		-0.25***		-0.24***
		(-4.62)		(-6.39)
HV-IV		0.28***		0.28***
		(4.55)		(8.26)
ILLIQ		0.08		0.02
		(1.60)		(0.53)
MFIS		-0.11***		0.16***
		(-3.84)		(5.81)
MFIK		0.01		0.14***
		(0.25)		(5.08)
VOIV		-0.07*		-0.07**
		(-1.89)		(-2.28)
Adj.R^2	0.26	4.94	0.23	4.95
Obs	84436	83936	84436	83936

Table 5
Option Portfolios Sorted by Different ML Textual Predictors

This table reports average monthly excess returns of the delta-hedged option portfolios sorted by machine learning (ML) textual predictors trained by alternative machine learning algorithms. The row labeled "SVR", "ENET", "RF", and "NN" reports portfolio sorting results by textual predictors extracted based on support vector regression, elastic net, random forest, or 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-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	-0.31	-0.14	-0.02	0.05	0.18	0.49***
		(-2.56)	(-1.07)	(-0.13)	(0.44)	(1.30)	(5.36)
	ENET	-0.27	-0.16	-0.05	0.04	0.21	0.48***
		(-2.31)	(-1.34)	(-0.34)	(0.32)	(1.39)	(4.84)
	RF	-0.29	-0.05	-0.04	-0.01	0.16	0.45***
		(-2.54)	(-0.37)	(-0.31)	(-0.05)	(1.10)	(5.15)
	NN	-0.29	-0.10	-0.06	0.05	0.16	0.45***
		(-2.40)	(-0.80)	(-0.45)	(0.42)	(1.05)	(4.57)
		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 − 1)**
Put	SVR	-0.56	-0.39	-0.35	-0.32	-0.23	0.33***
		(-6.34)	(-4.10)	(-3.70)	(-3.60)	(-2.35)	(4.92)
	ENET	-0.57	-0.39	-0.41	-0.30	-0.24	0.36***
		(-6.58)	(-4.38)	(-4.47)	(-2.96)	(-2.14)	(5.47)
	RF	-0.56	-0.41	-0.29	-0.33	-0.26	0.29***
		(-6.12)	(-4.63)	(-2.95)	(-3.33)	(-3.07)	(4.99)
	NN	-0.54	-0.39	-0.32	-0.34	-0.26	0.28***
		(-6.42)	(-4.17)	(-3.41)	(-3.54)	(-2.64)	(4.10)

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 ML textual predictors trained by using alternative word feature constructions. The row labeled "Unigram", "Bigram", "Trigram", or "Unigram + Bigram" reports portfolio sorting results based on ML textual predictors extracted based on different word features to train the model, including unigrams, bigrams, trigrams, or the combination of unigrams and bigrams, respectively. Detailed descriptions of these predictors are provided in Section 3.2.2. All returns are expressed in percentage. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey-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	-0.31	-0.14	-0.02	0.05	0.18	0.49***
		(-2.56)	(-1.07)	(-0.13)	(0.44)	(1.30)	(5.36)
	Bigram	-0.31	-0.16	0.00	0.06	0.16	0.47***
		(-2.49)	(-1.28)	(0.02)	(0.45)	(1.22)	(4.74)
	Trigram	-0.29	-0.08	-0.02	0.00	0.16	0.45***
		(-2.57)	(-0.65)	(-0.19)	(0.02)	(1.17)	(5.28)
	Unigram + Bigram	-0.31	-0.13	-0.04	0.04	0.22	0.53***
		(-2.45)	(-1.07)	(-0.32)	(0.29)	(1.53)	(5.14)
		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 − 1)**
Put	Unigram	-0.56	-0.39	-0.35	-0.32	-0.23	0.33***
		(-6.34)	(-4.10)	(-3.70)	(-3.60)	(-2.35)	(4.92)
	Bigram	-0.55	-0.41	-0.37	-0.29	-0.24	0.31***
		(-6.34)	(-4.06)	(-3.99)	(-3.23)	(-2.35)	(4.56)
	Trigram	-0.53	-0.33	-0.35	-0.36	-0.28	0.25***
		(-6.04)	(-3.42)	(-3.81)	(-4.15)	(-2.89)	(3.94)
	Unigram + Bigram	-0.58	-0.42	-0.33	-0.29	-0.23	0.34***
		(-6.56)	(-4.70)	(-3.63)	(-3.04)	(-2.39)	(5.22)

Table 7
Option Portfolios Sorted by Textual Predictors in Different Market Conditions

This table 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) illiquidity measure of individual stocks as a proxy for the market-level liquidity. "High Sentiment" ("Low Sentiment") is defined as periods with sentiment index higher than the median sentiment index for the entire sample period. Different periods regarding market volatility or liquidity are defined similarly. 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-West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Subperiod Analysis for Call Options						
High Sentiment	-0.19	-0.01	0.21	0.24	0.35	0.54***
	(-1.08)	(-0.06)	(1.02)	(1.36)	(1.70)	(3.98)
Low Sentiment	-0.43	-0.26	-0.24	-0.12	0.02	0.45***
	(-3.07)	(-1.94)	(-1.52)	(-0.82)	(0.16)	(3.57)
High VIX	-0.34	-0.15	0.06	0.04	0.23	0.57***
	(-1.55)	(-0.66)	(0.24)	(0.16)	(0.91)	(3.37)
Low VIX	-0.29	-0.13	-0.09	0.07	0.13	0.42***
	(-3.26)	(-1.28)	(-1.07)	(0.89)	(1.44)	(5.61)
High Liquidity	-0.22	-0.04	0.06	0.11	0.13	0.35***
	(-1.58)	(-0.30)	(0.31)	(0.77)	(0.89)	(3.85)
Low Liquidity	-0.43	-0.27	-0.13	-0.02	0.25	0.68***
	(-2.17)	(-1.15)	(-0.57)	(-0.11)	(0.96)	(4.06)
NBER Expansion	-0.30	-0.15	-0.02	0.05	0.12	0.42***
	(-2.52)	(-1.18)	(-0.16)	(0.45)	(1.01)	(4.99)
NBER Recession	-0.38	-0.03	0.01	0.06	0.78	1.16**
	(-0.73)	(-0.05)	(0.02)	(0.10)	(0.86)	(2.68)
Panel B: Subperiod Analysis for Put Options						
High Sentiment	-0.50	-0.26	-0.27	-0.13	-0.13	0.37***
	(-3.41)	(-1.75)	(-1.77)	(-0.95)	(-0.81)	(3.74)
Low Sentiment	-0.63	-0.52	-0.42	-0.49	-0.33	0.30***
	(-6.25)	(-5.27)	(-4.29)	(-5.30)	(-3.21)	(3.32)
High VIX	-0.68	-0.52	-0.45	-0.35	-0.24	0.44***
	(-5.03)	(-3.38)	(-3.02)	(-2.38)	(-1.41)	(4.12)
Low VIX	-0.45	-0.28	-0.25	-0.29	-0.22	0.23***
	(-4.97)	(-2.85)	(-2.88)	(-3.55)	(-2.78)	(3.04)
High Liquidity	-0.43	-0.25	-0.28	-0.31	-0.20	0.23***
	(-4.09)	(-2.23)	(-2.50)	(-3.77)	(-1.91)	(3.05)
Low Liquidity	-0.74	-0.59	-0.43	-0.32	-0.27	0.47***
	(-5.04)	(-3.56)	(-2.82)	(-1.92)	(-1.49)	(4.06)
NBER Expansion	-0.53	-0.38	-0.33	-0.33	-0.23	0.29***
	(-5.71)	(-3.74)	(-3.47)	(-3.85)	(-2.42)	(4.30)
NBER Recession	-0.91	-0.58	-0.53	-0.22	-0.19	0.72***
	(-3.03)	(-1.78)	(-1.64)	(-0.49)	(-0.45)	(3.62)

Table 8
Overlap Analysis of Dictionaries

This table 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 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 4.2.1. Figures in this table are expressed as percentage.

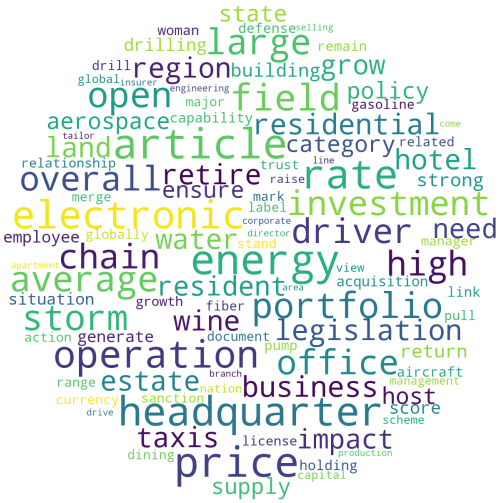
	Call (%)	**Put (%)**
ΔIV	19.85	18.70
IVOL	17.50	16.60
ILLIQ	8.55	9.65
MFIS	12.65	8.30
MFIK	11.00	9.10
VOIV	10.05	9.90
HV-IV	1.15	0.10

Table 9
Decomposition of Textual Predictors

This table reports the results of the regression analysis that decomposes the SVR option return 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 7. Independent variables are standardized to have 0 mean and unit standard deviation. The coefficients in this table are multiplied by 100. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey-West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

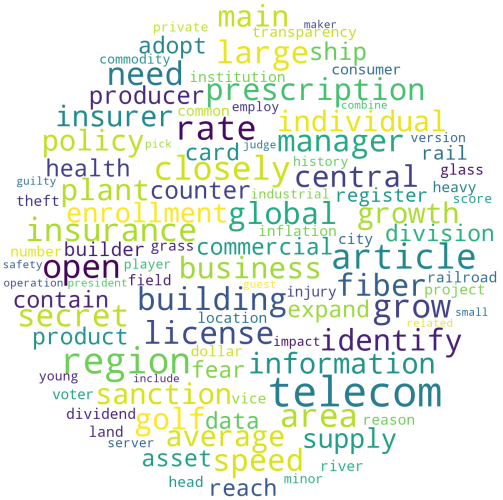
Panel A: Decomposition of *Call_SVR*								
ΔIV	0.10***							0.10***
	(12.93)							(13.86)
IVOL		-0.06***						-0.03***
		(-6.54)						(-7.49)
ILLIQ			-0.05***					-0.02**
			(-4.71)					(-2.48)
MFIS				-0.08***				-0.05***
				(-7.58)				(-6.09)
MFIK					0.03***			-0.01***
					(5.70)			(-3.48)
VOIV						-0.03***		-0.03***
						(-3.02)		(-6.10)
VRP							-0.01***	-0.01***
							(-2.68)	(-4.71)
Adj.R^2	16.19	9.27	9.22	11.83	2.68	5.92	2.73	38.14
obs	83936	83936	83936	83936	83936	83936	83936	83936
Panel B: Decomposition of *Put_SVR*								
ΔIV	0.09***							0.08***
	(11.70)							(13.55)
IVOL		-0.05***						-0.03***
		(-7.66)						(-5.70)
ILLIQ			-0.05***					-0.03***
			(-4.11)					(-3.56)
MFIS				-0.05***				-0.03***
				(-4.76)				(-3.27)
MFIK					0.02***			0.00
					(3.83)			(0.29)
VOIV						-0.04***		-0.05***
						(-5.74)		(-9.34)
VRP							-0.00	-0.01**
							(-0.65)	(-2.21)
Adj.R^2	11.59	6.91	9.36	8.18	2.49	5.08	2.19	30.67
obs	83936	83936	83936	83936	83936	83936	83936	83936

Figure 1. Word Cloud of Option Return Dictionaries

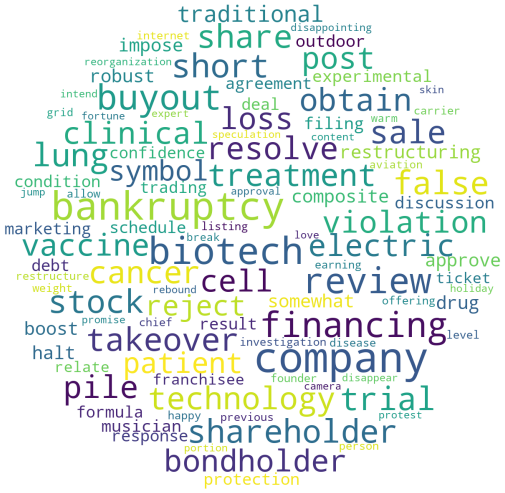


(a) Call Positive

(b) Call Negative



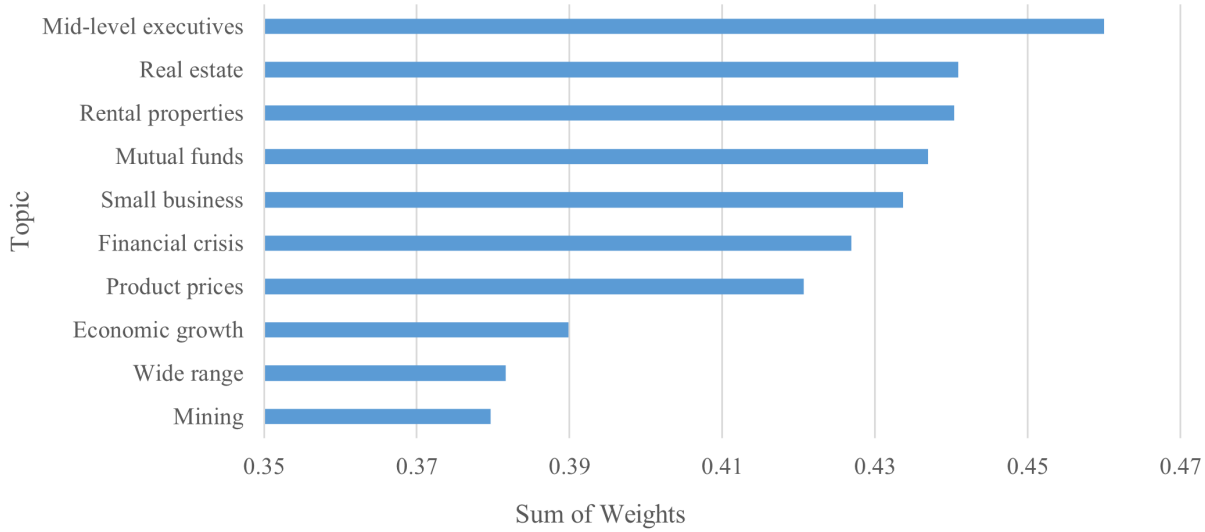
(c) Put Positive



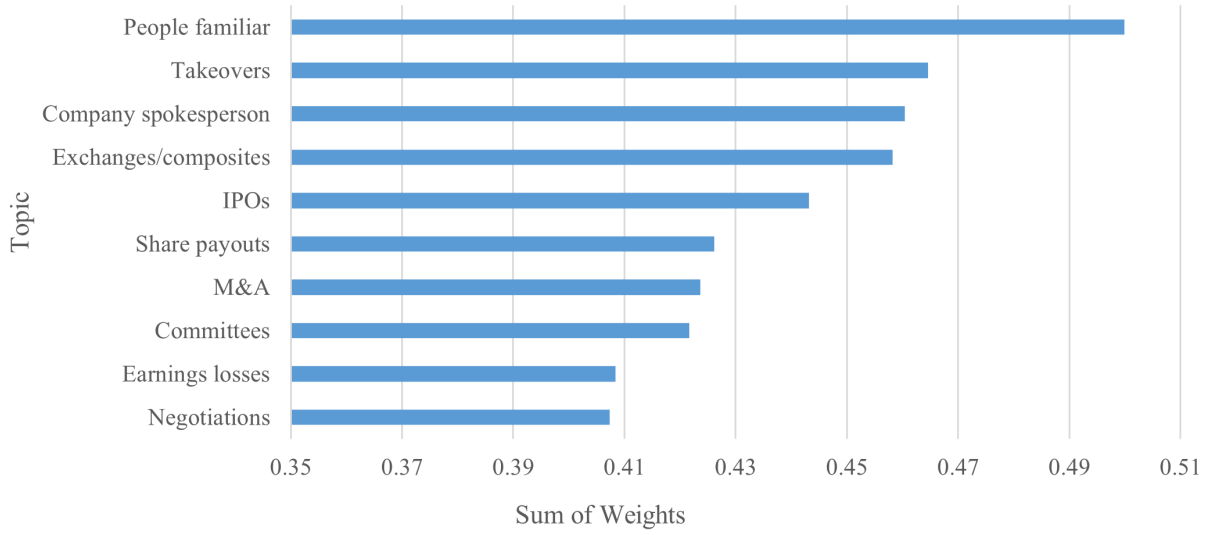
(d) Put Negative

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.

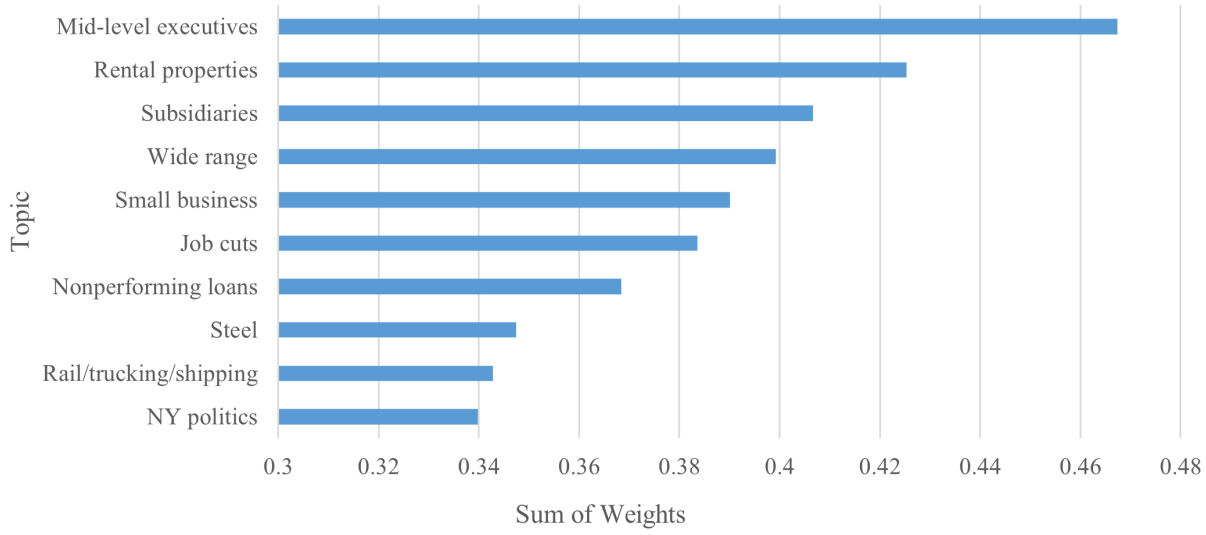
Figure 2. Top 10 Topics for Option Return Dictionary



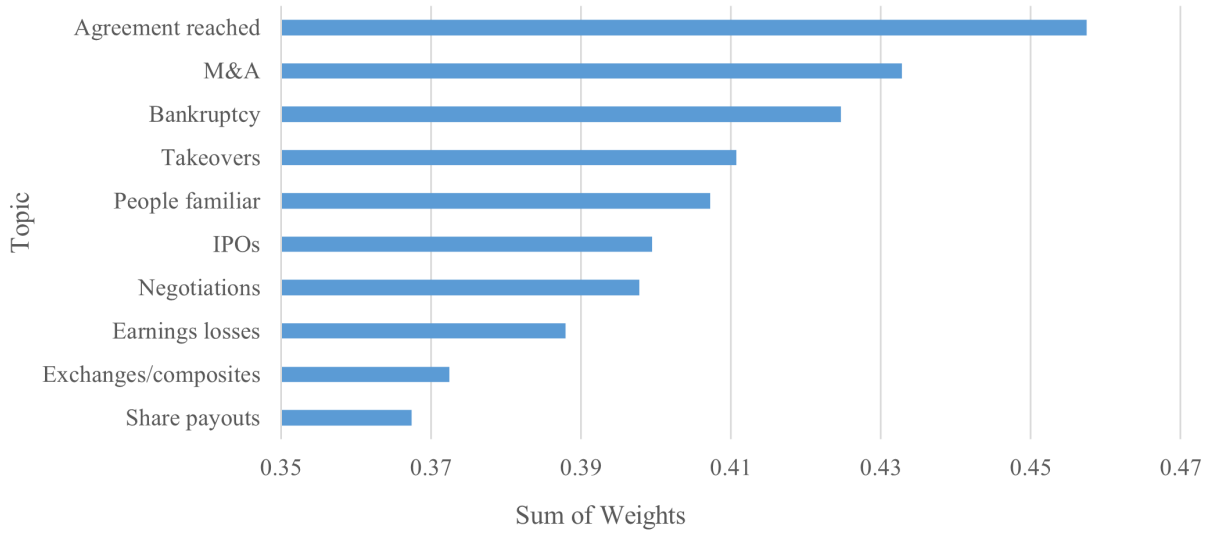
(a) Call Positive



(b) Call Negative



(c) Put Positive



(d) Put Negative

Figure 2: This figure displays the top 10 topics identified in each option return dictionary.

Figure 3. Causal Diagram for the Predictability of Textual Information

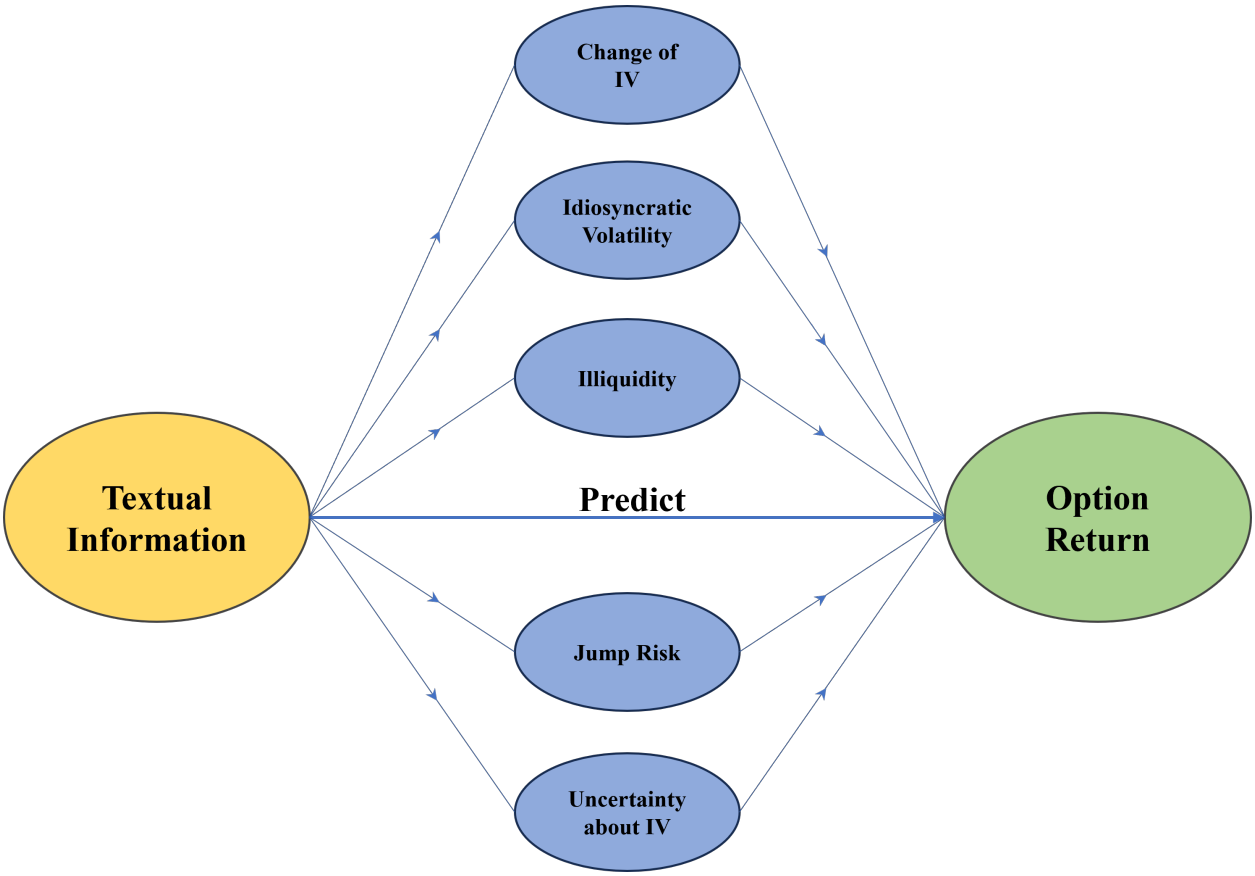


Figure 3: This figure illustrates the underlying logic of our economic mechanism analysis.

Figure 4. Illustration of the Calculation of Overlap Score

		Call Option		
		Positive	Non-Overlap	Negative
ΔIV	Positive	TRUE (328)	533	FALSE (139)
	Negative	FALSE (141)	510	TRUE (349)

(a) Call

		Put Option		
		Positive	Non-Overlap	Negative
ΔIV	Positive	TRUE (330)	534	FALSE (136)
	Negative	FALSE (150)	520	TRUE (330)

(b) Put

Figure 4: This figure shows how the overlap score is calculated between dictionaries for change of implied volatility and call option returns. $C = (328+349 = 677), N = (533+510 = 1043), W = (141+139 = 280)$, so that the overlap score is $(677-280)/(677+1043+280) = 0.1985$. 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
SVR	The textual predictors extracted from news media using the support vector regression model. SVR are estimated separately for call options (*Call_SVR*) and put options (*Put_SVR*). Textual predictors obtained using Elastic Net, Random Forest, or Neural Networks are called ENET, RF, or NN respectively.
	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 for call (put) options.
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 a Fama-French 12-industry distribution for the sample of stocks covered in our analysis and full CRSP sample. 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 Faburary 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)	3.64	1.10	1.68	3.14	3.89	4.40	4.88
Stock % coverage of stock universe (VW)	33.39	7.97	22.90	27.83	32.96	38.54	44.52
Stock % coverage of optionable stocks (EW)	9.38	4.19	3.09	6.29	9.47	13.06	14.77
Stock % coverage of optionable stocks (VW)	35.93	9.28	23.67	29.15	35.32	42.31	49.19
Stock % traded at NYSE/AMEX	69.68	5.12	61.02	67.13	70.71	73.11	75.40
Stock % included in S&P500 index	60.81	7.74	54.16	56.88	59.42	62.96	67.19
Stock % already included in last month	48.03	5.57	40.14	44.67	48.68	52.00	54.55

Panel B: Time-Series Average of Cross-Sectional Distributions (88,630 Stock-Month Obs.)

January 1996–November	Mean	Standard Deviation	10th Percentile	Lower Quartile	Median	Upper Quartile	90th Percentile
Firm Size in billions	32.63	68.60	1.37	3.31	10.65	30.82	79.21
Firm size CRSP percentile (%)	87.01	10.65	71.68	82.74	90.99	94.70	96.13
Firm book-to-market CRSP percentile (%)	34.41	25.02	5.11	13.18	29.49	52.83	72.70
Institutional Ownership (%)	65.03	17.20	41.38	56.72	68.44	77.35	83.35
Analyst Coverage	12.39	6.60	4.61	7.43	11.49	16.22	21.37

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

FF-12 Industry	Option sample	CRSP sample	FF-12 Industry	Option Sample	CRSP sample
Consumer nondurables	8.27	4.70	Telecom	2.77	2.66
Consumer durables	2.84	2.21	Utilities	2.54	2.25
Manufacturing	9.68	8.99	Wholesale	13.91	8.60
Energy	3.35	3.76	Healthcare	7.65	12.48
Chemicals	3.27	2.06	Finance	14.29	20.89
Business Equipment	19.73	17.77	Others	11.76	13.63

Table A2
N-gram Coverage

This table reports the percentage of token frequencies in the entire corpus that are accounted for by the top N most frequent tokens.

N	Unigram	Bigram	Trigram
1,000	62.29%	4.49%	1.08%
2,000	76.34%	6.13%	1.42%
3,000	83.12%	7.33%	1.66%
4,000	87.15%	8.30%	1.84%
5,000	89.84%	9.14%	2.00%
6,000	91.76%	9.88%	2.14%
7,000	93.19%	10.54%	2.27%
8,000	94.30%	11.14%	2.38%
9,000	95.18%	11.70%	2.49%
10,000	95.89%	12.23%	2.59%
40,000	99.89%	21.34%	4.25%
80,000	100.00%	27.66%	5.41%

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 ML 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-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	0.161***	0.162***	0.097***	0.110***	0.109***	0.058***
	(5.19)	(5.13)	(3.42)	(5.22)	(5.08)	(2.79)
LM_Sentiment	0.012		0.018	0.005		0.004
	(0.46)		(0.65)	(0.25)		(0.20)
LM_Uncertainty		0.048**	0.070***		0.001	0.005
		(2.07)	(2.67)		(0.07)	(0.28)
IVOL			-0.249***			-0.246***
			(-4.51)			(-6.41)
HV-IV			0.279***			0.280***
			(4.61)			(8.23)
ILLIQ			0.080			0.017
			(1.61)			(0.49)
MFIS			-0.114***			0.165***
			(-3.77)			(5.91)
MFIK			0.006			0.140***
			(0.18)			(5.08)
VOIV			-0.074**			-0.070**
			(-1.97)			(-2.32)
Adj.R2	0.400	0.341	5.157	0.271	0.358	5.051
Obs	84436	84436	83936	84436	84436	83936

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

This table presents the average monthly excess returns of delta-hedged option portfolios sorted by textual predictors developed using SVR but different number of word features. The rows labeled "4000," "6000," and "8000" correspond to the portfolio sorting results using textual predictors derived from the tf-idf matrix comprising the most frequent 4000, 6000, and 8000 words, respectively. The sample period is from January 1996 to November 2022. To adjust for serial correlations, robust Newey-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	4000	-0.25	-0.13	-0.06	0.05	0.16	0.41***
		(-1.97)	(-1.00)	(-0.52)	(0.41)	(0.99)	(3.95)
	6000	-0.24	-0.18	-0.00	0.03	0.15	0.39***
		(-1.98)	(-1.42)	(-0.02)	(0.22)	(1.03)	(3.86)
	8000	-0.29	-0.07	-0.07	0.08	0.12	0.41***
		(-2.30)	(-0.59)	(-0.60)	(0.54)	(0.85)	(4.39)
		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 – 1)**
Put	4000	-0.53	-0.45	-0.33	-0.31	-0.24	0.30***
		(-5.84)	(-5.11)	(-3.76)	(-3.20)	(-2.43)	(4.56)
	6000	-0.51	-0.40	-0.36	-0.35	-0.23	0.28***
		(-5.97)	(-4.21)	(-3.99)	(-3.67)	(-2.41)	(4.21)
	8000	-0.50	-0.38	-0.38	-0.36	-0.22	0.28***
		(-5.71)	(-3.91)	(-4.47)	(-3.83)	(-2.31)	(4.18)

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

This table 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 "3-month," "9-month," and "12-month" correspond to the portfolio sorting results when rolling window is 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-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	3-month	-0.30	-0.16	0.07	0.03	0.14	0.44***
		(-2.30)	(-1.28)	(0.47)	(0.28)	(0.93)	(4.36)
	9-month	-0.27	-0.09	-0.08	0.04	0.14	0.41***
		(-2.13)	(-0.62)	(-0.69)	(0.29)	(0.99)	(4.46)
	12-month	-0.32	-0.10	-0.03	0.09	0.14	0.46***
		(-2.62)	(-0.74)	(-0.27)	(0.65)	(1.00)	(4.98)
		1 (Low)	**2**	**3**	**4**	**5 (High)**	**(5 – 1)**
Put	3-month	-0.50	-0.43	-0.37	-0.34	-0.21	0.29***
		(-5.77)	(-4.46)	(-3.91)	(-3.72)	(-2.34)	(4.34)
	9-month	-0.52	-0.44	-0.32	-0.31	-0.27	0.25***
		(-5.87)	(-4.78)	(-3.06)	(-3.64)	(-2.76)	(3.82)
	12-month	-0.54	-0.42	-0.34	-0.27	-0.26	0.28***
		(-5.83)	(-4.72)	(-3.46)	(-2.97)	(-2.65)	(4.41)

Table A6
Correlations among Textual Predictors based on different Machine-learning Algorithms

This table reports the time-series average of 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, and neural networks.

		SVR	ENET	RF	NN
Call	SVR	1	0.72	0.60	0.52
	ENET		1	0.66	0.62
	RF			1	0.60
	NN				1
		SVR	ENET	RF	NN
Put	SVR	1	0.56	0.36	0.44
	ENET		1	0.52	0.67
	RF			1	0.43
	NN				1

Table A7
Option Return Dictionaries with Unigrams

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

	Positive	Negative
Call	energy, article, price, rate, electronic, operation, field, average, portfolio, storm, open, large, driver, office, overall, chain, investment, high, legislation, residential, business, wine, region, hotel, land, retire, taxis, water, estate, grow, impact, resident, need, category, policy, aerospace, state, ensure, host, supply, building, strong, return, drilling, score, employee, generate, situation, acquisition, link, remain, woman, aircraft, pump, holding, defense, capability, relationship, license, currency, related, trust, growth, manager, stand, mark, action, document, major, fiber, sanction, range, drill, globally, label, pull, merge, dining, gasoline, management, global, raise, scheme, nation, view, capital, engineering, corporate, tailor, come, director, drive, line, branch, area, insurer, apartment, production, selling, hold	trial, bankruptcy, potentially, somewhat, widespread, inaugural, clock, aviation, body, dilute, messy, breach, temporarily, clinical, craft, waiver, attempt, occasionally, indoor, left, depression, slack, survival, identification, supportive, consent, mask, accomplish, flier, approval, review, oxygen, console, schedule, concession, gamer, forbid, skip, flexibility, anxiety, glove, renegotiate, reporter, commissioner, resignation, predecessor, rumor, unemployed, sequence, bundle, postpone, game, passenger, symptom, videogame, brain, cooperate, upgrade, engage, fluctuate, broadband, patient, sick, arise, execute, prediction, treatment, creator, humor, restructuring, obligation, lesson, acute, burger, passionate, penny, loyal, setback, voting, disk, bury, performer, strap, publishing, subscriber, treat, disappointing, biotech, disease, negative, confirmation, hockey, complication, personally, workforce, burden, interface, virtual, fixing, bakery
Put	region, article, rate, area, open, grow, closely, need, insurance, license, building, large, secret, speed, global, fiber, manager, prescription, golf, main, plant, identify, central, business, growth, sanction, policy, information, individual, average, insurer, enrollment, supply, producer, ship, counter, health, asset, commercial, reach, expand, adopt, fear, product, data, contain, division, card, register, rail, builder, transparency, project, location, river, city, glass, head, institution, reason, heavy, vice, grass, inflation, dollar, theft, railroad, number, land, field, common, dividend, consumer, voter, injury, commodity, private, server, employ, history, minor, score, impact, version, young, industrial, player, operation, safety, have, include, guest, small, related, pick, judge, president, maker, combine, guilty	bankruptcy, workout, financing, false, break, company, dubious, endanger, achievement, knife, blade, knit, sentiment, practical, rebuff, math, insulation, closet, naturally, reject, violation, review, replay, talented, unlimited, lately, identical, dive, positively, spill, renegotiate, corruption, viable, survival, formula, criticise, buyout, grid, short, warm, resolve, symbol, aviation, nightmare, confidence, throw, dealer, mini, hardly, phase, twist, greenhouse, slot, customize, oust, flier, reporter, dominant, revise, defy, outsourcing, advocate, upbeat, tracker, lung, traditional, restructuring, experimental, outdoor, tone, manipulation, reorganization, willing, seller, homeowner, genome, potentially, instant, poise, gadget, appreciation, doughnut, loser, sequel, corn, creature, fluctuate, mechanic, method, severely, mode, theory, reception, ugly, bode, resonate, accomplish, dedication, premiere, initiate

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	area include, executive executive, insurance industry, hard come, practice know, average time, asset management, boost profit, focus group, market grow, supply chain, bank deposit, account bank, fight hard, senior vice, index fund, provide well, economic policy, grow share, tell customer, credit card, question people, bank financial, real estate, consumer good, public sector, plan time, number business, come close, grow demand, metropolitan area, card debt, share lead, base investment, invest fund, cost price, give chance, accord market, good time, large maker, take company, door open, multinational company, company serve, capital investment, middle ground, people money, estate investment, mutual fund, drive price, company security, know exactly, high profile, price average, system allow, large city, investment management, percentage point, investment vehicle, company asset, medium size, work force, customer come, double digit, double size, total return, early stage, long term, large firm, general manager, thing good, exploration production, report record, talk host, campaign finance, head home, book write, official decline, company eventually, company typically, emerge market, interest rate, rise rate, number high, large volume, step right, large market, pass legislation, public safety, long delay, late report, swimming pool, fund industry, people kill, cost time, great time, information available, industry group, make world, group executive	clinical trial, increase revenue, company result, send share, compare share, share share, patient receive, revenue earning, people line, seek approval, mortgage rate, company record, company statement, total company, share rise, company share, want bring, share revenue, develop technology, double number, fail meet, stock rise, hard drive, cash stock, drug market, company ability, forecast earning, difficult time, earning report, relate company, recent history, datum include, division include, drug company, hostile takeover, share accord, share cash, previously report, general counsel, take step, accord analyst, recently start, report profit, think twice, share period, large carrier, cost airline, drug approve, overseas market, company stock, company release, shareholder vote, work firm, drug treat, founder chief, company point, develop drug, company debt, meet requirement, company meeting, matter company, deal sign, meeting executive, deal company, time place, deal value, income rise, easy access, time make, generic drug, flight attendant, government affair, suit file, team people, revenue compare, play lead, question come, company face, share analyst, cancer treatment, revenue expect, senior adviser, share jump, accord people, send stock, domestic international, work help, research center, process take, deal likely, term long, thing learn, term deal, company develop, bankruptcy protection, revenue fall, hardware software, power generation, great opportunity, share information

Put	grow demand, senior vice, domestic market, comic book, meet need, safety issue, uphill battle, software developer, group plan, world company, asset management, market power, business employee, service revenue, hire worker, large city, continue build, medium sized, company stop, company typically, leave join, anytime soon, break ground, help shape, bank include, operate system, plan build, cause cancer, take responsibility, senior manager, estate firm, accord document, hand hold, hard sell, company head, file complaint, company investment, late report, entire life, significant impact, close home, project include, number high, officer company, long term, right activist, provide well, make process, product come, company operation, work thing, revenue come, business area, sentence prison, market base, time period, help fuel, make sense, vice president, face face, public affair, increase total, bank financial, civil liberty, think people, shopping center, take account, selling price, grow market, large maker, young woman, service launch, effort reach, knock door, know need, social issue, range issue, boost profit, consumer product, legislation require, well company, medium term, start small, percentage point, people help, work begin, computer chip, health plan, want hear, include work, bring people, member company, member tell, plead guilty, company corporate, long establish, government employee, accord local, business want, strong position	price jump, company statement, total company, company meeting, price share, estimate share, face company, company stock, company result, product sale, stay ahead, block deal, file bankruptcy, bankruptcy protection, company willing, help bring, meeting company, rise trading, bankruptcy court, share accord, company deal, sell store, company ability, stock fall, official believe, clinical trial, hold share, share climb, income rise, share rise, people high, come surprise, company plan, company debt, people place, stock rise, plaintiff lawyer, hard time, company share, domestic international, share jump, joint statement, time sale, begin career, company record, store company, role company, expect report, report decline, market make, question come, fail disclose, fall profit, change plan, develop drug, time give, place high, take helm, close company, deal fall, company analyst, instead take, play important, share sale, need additional, bond sale, post loss, term debt, investor confidence, cancer drug, move fast, accord people, patient receive, think great, overseas market, time base, reason believe, crash kill, trading stock, gain share, develop world, deal people, market stock, deal likely, include fund, share repurchase, want company, public company, hostile takeover, bear market, rating agency, earning company, start offer, conflict interest, good seller, medium company, willing accept, well know, general counsel, grow time

Table A9
Dictionaries for Option Return Determinants

This table lists the word feature importance for various option return determinants from the support vector regression (SVR) model. Option return determinants are the same with those mentioned in Table 8. Feature importance is defined as the top 1000 words with the largest magnitudes (i.e., the absolute value of the coefficients) from the SVR model. For abbreviation, we list the first 100 words which appear most often over time for both call and put options. The sample period is from January 1996 to November 2022.

	Positive	Negative
ΔIV	growth, revenue, rise, reach, strong, gross, average, movie, feed, segment, exclude, record, lift, period, gain, high, sale, result, litigation, final, manufacturer, price, complaint, income, rival, increase, rate, dollar, computer, home, analyst, housing, earning, double, transport, course, employee, valuation, video, profit, strengthen, employ, presence, report, agriculture, comparable, corn, compare, subpoena, footwear, laptop, system, pick, antitrust, interesting, ethanol, capture, coffee, beat, weigh, deliver, crop, hurt, name, energy, difference, score, stay, database, scan, main, year, single, prefer, manufacturing, release, metal, share, screen, album, company, return, truck, boost, refiner, house, margin, plaintiff, loss, exclusively, twice, teen, weakness, clothing, pair, association, compete, school, residential, insurance	coating, overture, conscious, schedule, financier, conspire, outflow, secure, accomplish, deployment, subcommittee, identification, respect, infant, sufficient, purchasing, routine, simulation, artistic, waive, bail, quick, automobile, tracker, stability, postpone, complexity, commitment, relation, reassure, uncertainty, stabilize, renegotiate, scenario, sister, bulk, persist, suitor, shake, prolonged, happy, powerhouse, victory, annualize, contraction, kidney, strip, tragedy, conserve, sizable, determined, coincide, breakfast, precedent, talent, stockpile, exposure, repay, outside, hardware, frozen, hammer, incremental, pill, permission, volume, foreclosure, rumor, contact, solution, considerably, taste, crush, patron, technique, ally, scrutinize, outline, dramatically, procurement, brief, object, southern, legislator, cast, tackle, poison, turmoil, marketer, greenhouse, gauge, soda, sentiment, tweak, legislature, acute, extent, appropriate, mission, gathering
IVOL	share, stock, trading, loss, plunge, news, tumble, online, company, drop, surge, fall, short, revenue, founder, rental, announce, disappointing, close, credit, soar, cash, trial, patient, closing, airport, biotech, genetic, gene, airline, sharply, chief, announcement, clinical, internet, music, mining, user, send, forecast, plummet, flier, inventory, struggle, site, mail, incorrectly, attendant, search, website, passenger, analyst, gross, flash, rally, peace, halt, liquidity, bankruptcy, tech, downgrade, found, lower, disease, seller, cell, chip, computer, capitalist, commute, jump, hope, unprofitable, clothing, royalty, treatment, memory, debt, investigation, expectation, medium, gambling, storage, amortization, solar, luggage, steel, delay, base, drilling, destination, resign, flight, investor, ability, athletic, hedge, fetch, possible, enrollment	vice, career, packaging, unit, toothpaste, soap, reactor, railroad, partly, paint, smoothly, mutual, investment, global, division, diaper, currency, chocolate, catastrophe, multinational, snack, modest, candy, color, starter, appointment, underwriting, issuer, medication, instance, tender, cereal, recycling, sugar, graduate, tournament, beverage, freight, retirement, banking, corporate, sensor, frozen, coating, acquisition, nuclear, inventor, inflow, pulp, abroad, diversified, syrup, blade, kidney, vary, syndicate, rapper, commercial, innovation, head, insurance, consumer, boost, award, join, foreign, offset, appeal, spokesman, business, branch, wife, train, survey, responsibility, asset, buyback, green, signal, scandal, golf, successor, affair, remember, tissue, sensitive, broadcast, laundry, argue, wholesaler, sluggish, bureaucracy, jetliner, lineup, celebrate, flavor, alleviate, cocoa, awareness, nutrition

HV-IV	share, revenue, shareholder, premium, slightly, user, beat, well, acquisition, report, gross, equity, site, coach, software, rebound, maker, price, function, power, synergy, closing, customer, value, agree, computer, approval, compete, stock, downturn, offering, deal, firm, period, hospital, rise, identify, play, acquire, forecast, position, demand, technology, opportunity, settlement, prior, transaction, chip, suggest, pick, size, cent, major, smart, good, rent, offer, salary, release, enable, space, compare, combined, recipe, equipment, telecommunication, moderate, potentially, disease, come, suit, surge, summer, presence, double, expectation, cash, victim, administrator, common, combine, sign, cinema, institutional, friendly, information, head, partly, symptom, substantially, hefty, desktop, picture, meet, combination, list, player, card, subscription, takeover	await, crush, altogether, sweater, crimp, implant, multinational, cereal, sophisticated, sideline, deepen, downgrade, deteriorate, promotion, side, defensive, punitive, identical, reinsurance, rural, explode, coat, halftime, subprime, grim, foreclosure, flock, hurricane, doll, printing, frequently, scramble, turmoil, shutdown, worried, strap, locate, shield, absence, today, retrieve, pain, amount, prototype, undervalue, hurry, unusually, deterioration, proprietary, fingerprint, feed, quarterback, game, staff, effect, pressure, material, cede, reception, inability, load, motion, criticise, formally, taxpayer, confidential, wipe, enact, sour, wallet, jack, rumor, cotton, regime, continued, lately, illustrate, foster, fragmented, oxygen, fortune, wheel, airfare, fluid, tour, cease, curb, disaster, flight, commit, loose, formation, workforce, brake, chairman, associate, underperform, tumble, meat, professor
ILLIQ	loss, contemporary, debt, liquidity, playwright, dancer, dear, resign, photography, fiction, fare, firm, news, solo, choreographer, peace, closure, swim, chicken, guitar, airport, musician, adviser, rental, inmate, passenger, recreational, consolidation, hope, voter, steel, prisoner, plummet, piano, terrace, vehicle, murder, exhibition, gene, estate, faculty, boat, bookstore, composer, bureau, stone, collapse, outflow, short, publicly, survivor, staff, athletic, lose, choreography, inquiry, music, close, base, historian, literary, referendum, escape, search, artist, specialist, postwar, online, meat, intelligence, attendant, real, soprano, public, photograph, firearm, jewelry, recording, flight, musical, fireplace, vocal, publication, poultry, assessment, pension, mourn, nominee, diplomatic, ballet, writer, investigation, buyer, rent, film, ballot, independence, rescue, cash, villa	viewership, stent, spokesman, soda, soap, shampoo, server, router, razor, project, garbage, prescription, plaintiff, pipeline, pill, packaging, multinational, locomotive, license, innovation, guru, growth, giant, unit, valuable, earning, telecast, toothpaste, insurer, exploration, processing, drug, discriminate, diaper, detergent, customer, currency, conglomerate, computing, compute, chip, carry, career, data, cable, exclude, capability, enable, arthritis, behemoth, beverage, barrel, blockbuster, broadband, bundle, buyback, blade, daughter, variant, procurement, massive, dominant, jetliner, toilet, pulp, large, programmer, medical, technical, holding, dominate, guidance, manufacturer, fiber, important, sponsorship, bulk, swipe, rollout, regulation, smartphone, defibrillator, consensus, birdie, giving, cheaply, disk, remove, reactor, eclipse, cartridge, machinery, infect, reinsurance, name, monopoly, philanthropic, allergy, format, unveil

MFIS	download, miner, internet, audio, content, software, user, computer, online, digital, gambling, technology, video, loss, memory, search, airline, subscriber, gaming, mining, genetic, carrier, subscription, steel, delete, graphic, announcement, click, hacker, news, useful, company, panel, inventory, rumor, stream, browser, music, price, share, disappointing, fall, server, drop, wireless, vacation, venture, storage, attribute, exploit, desktop, antitrust, trading, producer, barrel, biotech, flier, casino, tech, lung, analyst, capacity, privacy, trial, tourist, passenger, ramp, brick, gene, gold, aluminum, blog, finally, mortar, travel, clinical, chat, console, virtual, revenue, airport, version, resign, compete, clothing, virus, site, access, frequent, game, platform, treatment, traveler, flash, extract, marketplace, screen, publisher, gadget, exploration	shampoo, crisis, banking, portfolio, eurozone, servicing, overrun, elevator, wrongdoing, economy, assume, dividend, bond, insurance, rate, unit, bank, client, area, default, teller, stadium, clearing, divestiture, oversight, cushion, adequate, subprime, landfill, insight, yuan, drain, conditioning, damage, punitive, participant, emerge, risky, institution, liquidity, risk, lender, lending, issuer, borrower, manage, insurer, fund, program, derivative, economist, capital, instrument, asset, industrial, denominate, indicator, beverage, consultant, benchmark, annualize, loosen, tractor, overall, warehouse, bulk, yard, reinsurance, antidepressant, upside, wealth, procedure, refinance, attorney, holding, utility, mutual, income, light, investment, retirement, borrowing, recession, interest, rank, blue, yield, aerospace, liability, deposit, operation, hold, banker, pension, coating, razor, prescription, foreclosure, issuance, scandal
MFIK	utility, bank, food, tobacco, cigarette, snack, rate, takeover, fund, brand, acquisition, retire, branch, dividend, unit, packaging, director, buyout, portfolio, asset, pension, electricity, increase, group, insurance, regulator, cereal, equity, line, beverage, income, banking, system, industrial, regulatory, structure, business, saving, division, agency, film, mutual, premium, aerospace, graduate, manager, rural, meat, insight, catastrophe, plant, come, drink, investment, change, power, offset, water, nuclear, spirit, yield, sugar, inning, election, smoker, regional, butter, lender, science, package, base, candy, inflation, survey, acquire, railroad, auction, underwriting, defense, freight, agree, category, chairman, currency, deal, bureau, chocolate, master, percentage, documentary, smoking, synergy, property, distributor, state, grow, earning, institution, paper, credit	software, portal, miner, memory, drilling, slot, stupid, skip, granddaughter, arguably, blog, reviewer, mileage, mortar, click, flier, stock, clothing, chip, delete, subsidy, optical, handmade, steep, hate, speaker, luggage, departure, brick, videogame, niece, patient, video, plunge, download, online, website, screen, semiconductor, internet, tech, attendant, accountable, symptom, drop, jean, touch, shale, cancellation, disappointing, surge, toss, plug, determination, steelmaker, automaker, spiral, infection, password, aluminium, booming, virus, server, cell, cabin, wait, button, downside, catalog, baggage, destination, gallon, month, extended, quickly, flagship, printer, cure, tennis, unveil, entrepreneur, success, guidance, jacket, voucher, vintage, objection, taxi, recipient, sedan, search, sick, leather, computer, booking, equipment, flight, absence, streaming, subscription

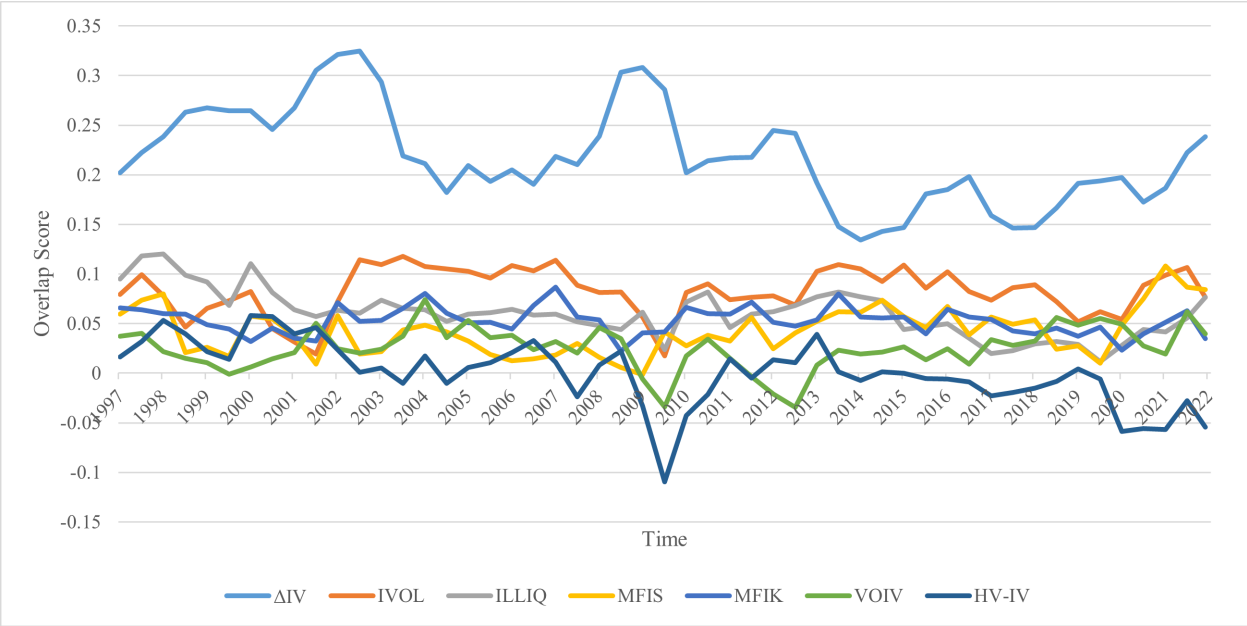
VOIV	deal, debt, takeover, share, credit, shareholder, announce, acquisition, cash, utility, regulator, value, transaction, company, offer, close, loan, trading, wave, stock, plunge, firm, disclosure, buyout, short, loss, premium, closing, purchase, tumble, reject, acquire, investor, consolidation, fall, filing, refinance, vote, volatile, news, bank, contribute, merger, market, meat, board, troubled, acquirer, receive, result, investigator, transmission, complete, failure, emergency, time, finish, shock, bankruptcy, base, term, cooperate, electric, suitor, federal, worsen, turmoil, saving, lender, regulatory, biotech, equity, investigation, sector, synergy, executive, approval, specialize, cause, selloff, branch, care, screening, trader, advisory, judge, fraud, leverage, risk, asset, significant, subpoena, depreciation, assume, electricity, previous, possible, affect, nation, volatility	chest, politic, runway, haul, barge, amenity, chocolate, scheduling, diaper, organize, interesting, warm, railroad, lifestyle, refining, site, tenant, traditionally, replicate, pillow, annualize, lifelong, button, videogame, renovate, guest, tanker, debate, constant, clothe, airplane, gallon, council, quarterback, outpace, rely, heating, convenience, embrace, wrap, racial, integrated, illustrate, working, optimism, idea, barrel, treasurer, cargo, ship, mainstream, refinery, gasoline, floor, crude, production, lure, wheat, self, conventional, realise, shopper, expenditure, choke, import, pink, society, disappointment, drift, calendar, wing, roster, mineral, border, understanding, designate, harassment, console, loyalty, movement, defer, objective, recruitment, comfort, general, demand, offshore, welcome, want, departure, pant, agenda, boot, attendant, experienced, seasonal, poverty, youth, cutter, applicant

Table A10
An Example of the News Article

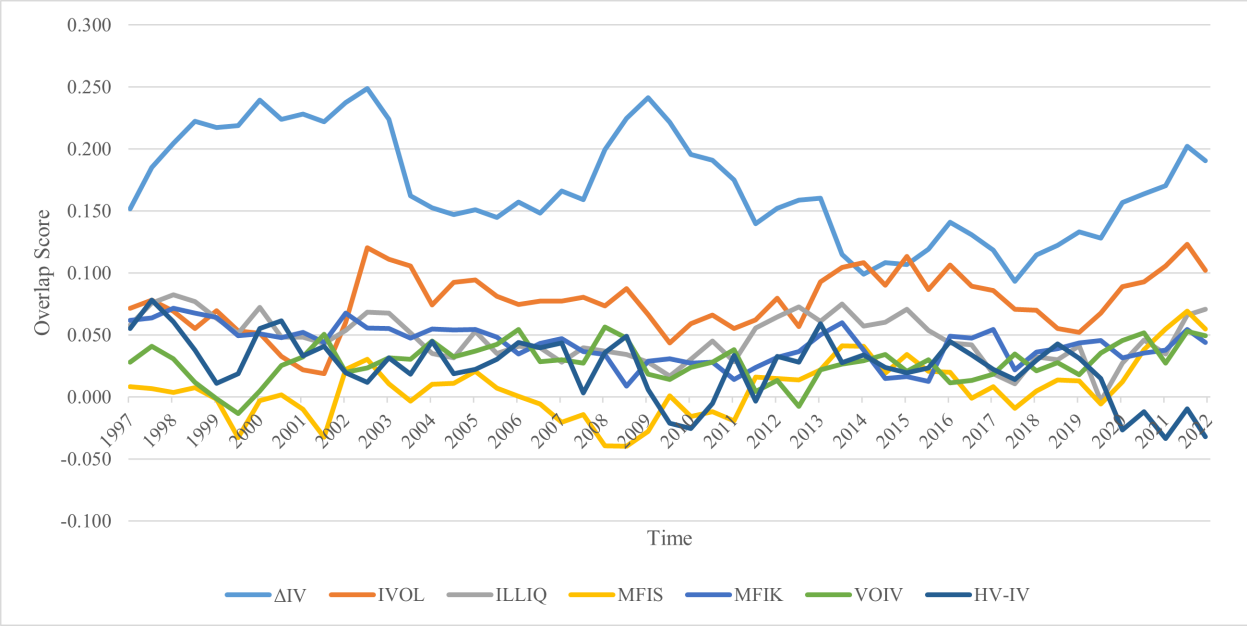
This table shows a news article included in our sample. As an example, we highlight words that positively (negatively) forecast delta-hedged call option returns with red (blue).

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.