Forecasting Option Returns with News[*]

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

[February 2023]

Abstract

This paper investigates whether news articles on companies contain useful information about future equity option returns. We apply machine-learning approaches to extract signals from text data that can significantly predict delta-hedged option returns. Our results are robust after controlling for known option return predictors. More than half of the important words are sentiment-related. We find evidence that one channel underlying the explanatory power of the textual predictors based on machine-learning approaches is their information content about change in future volatility. Our work highlights the importance of analyzing unstructured data like texts for pricing derivatives and provides new evidence for machine-learning approaches' superiority in extracting information from unstructured data.

Keywords: textual analysis, option return predictability, machine learning, investor sentiment, news media, alternative data

JEL classification: G12, G13, G14, G17

[*] We thank Turan Bali, Hector Chan, Evan Jo, Hugues Langlois, Asaf Manela, Paul Tetlock as well as seminar participants at CFE-CMStatistics (2021), CICF (2022), AsianFA (2022), and CIRF (2022). We acknowledge financial support of the Research Grant Council of the Hong Kong Special Administrative Region, China (Project No. GRF 14500919). All errors are our own.

[†] Jie Cao is at The Hong Kong Polytechnic University, jie.cao@polyu.edu.hk. 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. Bing Han is at University of Toronto, bing.han@rotman.utoronto.ca. Xintong Zhan is at Fudan University, xintongzhan@fudan.edu.cn.

1. Introduction

Unstructured data, such as texts, images, and videos, contain important information about firm fundamentals and stock performances. For example, 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 content can capture otherwise hard-to-quantify aspects of firms' fundamentals.[1] Some recent work starts exploring more advanced natural language processing tools to extract information from unstructured data. For example, Frankel, Jennings, and Lee (2022) document that machine-learning (ML) methods can produce more powerful and reliable disclosure sentiment measures than dictionary-based methods.[2]

Despite the rich application of text data to study the equity market, little is known about the applicability of textual analysis in the options market. In this paper, we fill the gap by extracting novel information from news media via machine-learning approaches that can be used to predict cross-sectional equity option returns. Such text-based option return predictability is robust and generates sizeable option trading profits. Furthermore, we document that machine-learning approaches capture information from various perspectives, such as investor sentiment and future change in volatility. Our results showcase the usefulness of text data and machine learning tools for a better understanding of the options market.

We start our investigation by training a support vector regression (SVR) model to learn the relation between texts and future option returns using over five million news articles. SVR is a supervised machine-learning algorithm that can effectively handle large dimensional feature space and has been shown to work well in finance (Manela and Moreira 2017).[3] Our SVR predictor based on news media coverage of companies significantly predicts equity option returns. When we sort options into quintile portfolios based the SVR predictor, the 5-minus-1 portfolio return spread is significantly positive averaging 1.48% (0.86%) per month for call (put) options and survives realistic option transaction costs. Such predictability holds for

[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. Hassan, Hollander, van Lent, and Tahoun (2019) and Engle, Giglio, Kelly, Lee, and Stroebel (2020) utilize textual analysis to measure firm-level political and climate change risks, respectively.

[2] From methodological perspectives, Ke, Kelly, and Xiu (2019) and Kelly, Manela, and Moreira (2021) improve the ability of machine-learning approaches to extract information from text data.

[3] Manela and Moreira (2017) construct a text-based measure of uncertainty using support vector regression and apply the measure to predicting VIX.

different option moneyness or maturity. Our results are robust under alternative machine-learning methods, including elastic net, random forest, and neural networks, and hold for different word constructions (such as bigram, trigram, and fivegram). Furthermore, the SVR predictor based on text data from newspaper is distinct from existing option return predictors such as volatility deviation (Goyal and Saretto 2009), idiosyncratic volatility (Cao and Han 2013), stock return autocorrelation (Jeon, Kan, and Li 2019), and various underlying stock characteristics (Zhan, Han, Cao, and Tong 2022).

We conduct further tests to understand the nature of SVR textual predictors for equity option returns. We find that the majority of the important word features selected by SVR comes from sentiment-related dictionaries. For example, more than half of the important words that contribute to the textual predictors come from two sentiment dictionaries widely used in the literature, namely the General Inquirer dictionary and the Loughran-McDonald dictionary. We then construct two sentiment-related measures based on these two dictionaries using traditional methods. One measure reflects the sentiment level, and the other reflects the sentiment disagreement. SVR predictors are closely related to lexicon-based sentiment measures as well as RavenPack sentiment measures. The predict power of SVR predictors for option returns is more significant (both economically and statistically) than lexicon-based sentiment measures and remains robust after controlling for the latter.

In addition to capturing sentiment-related information, we find that the SVR predictor is significantly positively related to the implied volatility change over the next month. In contrast, we do not find similar predictability for lexicon-based sentiment measures. We also find that the SVR textual predictor is not related to future stock returns or earnings surprises.

Our paper contributes to the growing literature of option return predictability. Recent studies have documented many predictors for equity option returns (see e.g., Zhan et al. (2022)). To the best of our knowledge, we are the first to forecast equity option returns using information from news media and machine learning approaches. We demonstrate that machine learning techniques are powerful in extracting important information embedded in the news media. Our paper is closely related to Bali, Beckmeyer, Moerke, and Weigert (2021), which applies machine learning techniques in forecasting option returns. Their predictors, stock characteristics and option-based characteristics, are distinct from the textual predictors we extract from news media. Our paper complements their study by demonstrating the benefit of machine learning techniques applied to alternative data that generate useful information for predicting future 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 the different economic channels and potential explanations of the equity option return predictability based on news media information. Section 5 concludes the paper.

2. Data and Sample

2.1. Data and Sample Descriptions

The newspaper data is mainly collected from ProQuest and complemented with Factiva. From February 1996 to December 2018, 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. Since most articles in ProQuest and Factiva do not have firm-specific tags, we need to identify and match each article to 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 firms' names 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 by 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 exclude those firms that are difficult to be identified by company names (e.g., including common words). Every month, we remove the outliers (firms with more than 50 articles) to minimize matching errors. Further, we manually go through a randomly selected subsample with all available company names from our matched article database to ensure each article is correctly assigned to affiliated companies by the fuzzy matching algorithm. Since we are interested in equity option returns, we only select firms that have both newspaper coverages and equity options traded. After all the filtering, we have 5,928,866 article-month observations in total. Compared with other studies, our dataset covers a broader scope of news media information for firms traded in the U.S. market.

For option return sample, we collect equity option data, including best bid, best offer, expiration date, and strike price from the OptionMetrics database. The underlying stock variables, such as stock return, stock price, trading volume, and shares outstanding, are collected from the CRSP database. Analyst forecast dispersion data is obtained from IBES. The sample period is from February 1996 to December 2018. For each month, we choose individual equity options that are closest to being at-the-money and have the shortest maturity among those with more than one month to expiration. Following the existing literature, we exclude observations that violate no-arbitrage conditions, have no trading volumes or open interests, have a quoted mid-price less than $0.125, and have paid cash dividends during the holding period. The holding period is from the beginning to the end of each month. We only retain stocks with both call and put options available after filtering.

Our final sample contains 50,888 option-month observations for both call and put options on individual stocks over the 275-month sample period. On average, we have 185 option observations for each month. Panel A and B of Table 1 show the summary statistics for our option return sample.

[Insert Table 1]

2.2. Variable Constructions

Our main independent variables are the machine-learning (ML) textual predictors since machine-learning approaches are well suited to deal with text data. First, we follow Manela and Moreira (2017) to construct a large set of potential information unigrams (with bigram noun phrases identified by the NLTK Python package) and select useful words among all text features. Different from the lexicon-based approach and Manela and Moreira (2017), the contents of our machine learning dictionary are time-varying, as news media typically changes attention and coverage over time. For example, from 2016 to 2020, "Donald Trump" may significantly impact the financial markets and receive more attention from the public, although things will be largely different before 2016 and after 2020 when the political news had less attention on "Donald Trump." Therefore, complementing the dictionaries of Tetlock (2007) and Loughran and McDonald (2011), we apply machine learning techniques to better capture the shift in the media focus by allowing words to be in and out of the important list over time.

Instead of transferring word counts to a counting matrix that assigns equal weights to all words, we adjust the count numbers by the process of *tf–idf* (term frequency-inverse document frequency) commonly used in natural language processing. For each word j in the word list, the *tf–idf* weighted value for article h about firm i at time t is defined as:

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

where $tf_{i,t}^{h,j}$ is the frequency of occurrence of the word j in the article, $w_t^{j,idf} = \log \frac{H_t}{df_t^j}$ with $H_t = \sum_{i=1}^{N_t} H_{i,t}$ defined as 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 in the sample. The adjusted weight for each feature in the word list is used as input to fit the support vector regression with our target variable, delta-hedged call/put option returns. For each firm i, the word feature j for firm i is given by taking the average among all articles covering firm i $(H_{i,t})$ at time t:

$$x_{i,t}^j \equiv w_{i,t}^j = \frac{1}{H_{i,t}} \sum_{h=1}^{H_{i,t}} w_{i,t}^{h,j,tfidf}. \tag{2}$$

Second, after obtaining the word features, we apply machine-learning techniques to the text data. In our study, the traditional statistical methods do not work well since the independent variables involve high-dimension data (e.g., more than 10,000 single words and dual-word combinations). In a seminal paper, Manela and Moreira (2017) apply the support vector regression to construct a news-based VIX through high-dimension 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:

$$r_{i,t} = \alpha_t + \beta_t' x_{i,t-1} + \epsilon_{i,t}, \quad i = 1, 2, \ldots, N_t, \tag{3}$$

where $r_{i,t}$ is the variable of interest, which is either call or put delta-hedged equity option returns, for firm i at time t. $x_{i,t-1} = \left[x_{i,t-1}^1, \ldots, x_{i,t-1}^K\right]'$ is a $K \times 1$ vector of (all the) K word features from the newspaper articles related to firm i at time $t-1$. We omit those words that appear less than three times in the entire sample. Since the linear regression above is with high dimensions (more than 10,000 words every month), the traditional OLS cannot be used to estimate β_t. To solve the problem above, we apply the support vector regression (SVR) approach, which can be formulated as:

$$\beta_t^* = \arg\min_{w} \frac{1}{2} \|\beta_t\|_2 + C \sum_{i=1}^{N_t} \left(\xi_{i,t} + \xi_{i,t}^*\right),$$

$$\text{subject to} \begin{cases} r_{i,t} - \beta_t{}' x_{i,t-1} - \alpha_t \leq \varepsilon + \xi_{i,t} \\ \beta_t' x_{i,t-1} + \alpha_t - r_{i,t} \leq \varepsilon + \xi_{i,t}^* \,, \\ \quad\quad \xi_{i,t}, \xi_{i,t}^* \quad \geq 0 \end{cases} \qquad i = 1, 2, \dots, N_t. \tag{4}$$

The intuition 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 points fall outside the ε margin. As such, we need to account for the possibility of errors that are larger than ε. Following Cortes and Vapnik (1995), we introduce slack variables $\xi_{i,t}, \xi_{i,t}^*$ to cope with 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 linear function and the amount up to which deviations larger than ε are tolerated. This corresponds to dealing with the so-called ε- insensitive loss function $|\xi|_\varepsilon$ described by:

$$|\xi|_\varepsilon := \begin{cases} 0, & \text{if } |\xi| \leq \varepsilon \\ |\xi| - \varepsilon, & \text{otherwise} \end{cases}. \tag{5}$$

The problem above can be solved in its dual formulation (see Schölkopf and Smola 2002). To construct a pure out-of-sample signal, at each point of time t, we use all available word features extracted from the newspaper and equity option returns in the past three months to train the model. The textual information proxy (i.e., textual predictor) for future equity option returns at time $t+1$ based on available news media at time t is then given by:

$$TP_{i,t} \equiv \hat{r}_{i,t+1} = \hat{\alpha}_t + \hat{\beta}_t' x_{i,t}, \qquad i = 1, 2, \dots, N_t. \tag{6}$$

This fitted value is used as the time t textual signal for the next period's equity option returns. We conduct various tests to evaluate the predictive performance of textual predictors for equity option returns. Using the rolling window to train the model, we also allow the dictionary (i.e., word features) to change over time. Our methodology can capture time variation in news media coverage and adapt to new textual information. When constructing textual predictors, we treat delta-hedged call and put option returns separately to train different models, although delta-hedged call and put option returns are highly correlated due to the put-call parity relationship.

Panel D of Table 1 reports the time-series average of the cross-sectional correlations between ML textual predictors and option or equity characteristics. Although ML textual predictors have relatively high correlations with each other, their correlations with option or equity characteristics are generally low.

In addition to the support vector regression, we also consider other machine learning methods, such as elastic net, random forest, and neural networks, to deal with the high-dimensional data of news media and capture potential nonlinearity and interactions among independent variables. We use SVR for our main empirical results because it is more interpretable and stable with fewer hyper-parameters to adjust, therefore less subjected to data snooping issues. We apply alternative machine learning approaches as robustness checks on our empirical results (see detailed model specifications and their performances in Section 3.2).

Our main dependent variable is the delta-hedged option returns. To eliminate the exposure to the underlying stock, we examine delta-hedged option returns.[4] Our buy-and-hold delta-hedged option return (HRP) equals the total dollar gain at the end of the holding period scaled by the absolute value of the total cost of constructing the portfolio at the formation date. Specifically, the delta-hedged call option return over $[t, t+1]$ is given by:

$$HRP_{i,t+1}^{C} = \frac{H_{i,t+1}^{C}}{H_{i,t}^{C}} = \frac{(C_{i,t+1} - C_{i,t}) - \Delta_{i,t}^{\mathrm{C}}(S_{i,t+1} - S_{i,t})}{\left|C_{i,t} - \Delta_{i,t}^{\mathrm{C}} \times S_{i,t}\right|}, \tag{7}$$

where the initial investment cost is $H_{i,t}^{C} = \left|C_{i,t} - \Delta_{i,t}^{\mathrm{C}} \times S_{i,t}\right|$, with *C* and *S* denoting the call option price and the underlying stock price. $\Delta_{i,t}^{\mathrm{C}}$ is the Black-Scholes call option delta for firm i at time t. The total dollar gain at the end of the holding period is $H_{i,t+1}^{C} = (C_{i,t+1} - C_{i,t}) - \Delta_{i,t}^{\mathrm{C}}(S_{i,t+1} - S_{i,t})$. Similarly, delta-hedged put option return for firm i is given by:

$$HRP_{i,t+1}^{P} = \frac{H_{i,t+1}^{P}}{H_{i,t}^{P}} = \frac{(P_{i,t+1} - P_{i,t}) - \Delta_{i,t}^{\mathrm{P}}(S_{i,t+1} - S_{i,t})}{\left|P_{i,t} - \Delta_{i,t}^{P} \times S_{i,t}\right|}, \tag{8}$$

note that put option's delta, $\Delta_{i,t}^{\mathrm{P}}$, is negative. Panel A (B) of Table 1 presents the summary statistics for delta-hedged call (put) option returns and call (put) option characteristics.

[4] Tian and Wu (2021) show that the monthly rebalanced delta-hedging strategy can remove about 70% 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. (2021).

3. Empirical Results

3.1. Baseline Results

3.1.1. Single Portfolio Sorts

To apply machine learning approaches to forecast equity option returns, we first use the support vector regression to fit the option returns with the textual information from the newspaper articles over the previous three months. Next, we refer to the predicted delta-hedged call or option returns over the next month by SVR based on the textual information over the previous three months as the textual predictors, denoted by Call_SVR and Put_SVR. We then sort firms based on textual predictors into quintiles and compare the next-month realized returns of the sorted portfolios.

Each month, we track what words have the most important contributions to explaining the target variables across firms in the training sample. The feature importance of a given word is defined as the absolute value of the corresponding coefficient from the support vector regression. We will list those words in the newspapers with the largest magnitude of the coefficients averaged over time. In Section 4, we will classify these important words into different groups/topics such as sentiment, politics, macroeconomics, and firm-specific. This exercise provides useful insights into the underlying mechanism of the option return predictability by the textual information extracted through support vector regression.

Table 2 shows that textual information provides significant and robust predictive power of delta-hedged equity option returns using various portfolio weighting schemes. The magnitude of the monthly long-short option strategy is economically significant and comparable to other option return predictors in the literature. For example, the option portfolio return spread between the bottom and the top sorted by the textual predictors using SVR can be as large as 1.48% (0.86%) on average per month for call (put) options. Following Eisdorfer, Goyal, Zhdanov, and Boulatov (2022), we also report the alphas after adjusting the delta-hedged option returns using a 7-factor model, and our results are robust to the risk adjustment. The seven-factor model includes five stock factors in Fama and French (2015), the momentum factor, and the option factor in Coval and Shumway (2001). Although the absolute value of option return shrinks for each portfolio, the 7-factor alpha of the spread portfolio is barely different from the average raw return, indicating our results are robust to risk adjustments.

[Insert Table 2]

In addition to conducting analysis assuming that options can be traded at their mid-price, we also consider profits after various amount of transaction costs. Compared with the stock market, the equity option market has significantly higher transaction costs. In our sample, stocks are typically large and liquid, with a relatively low bid-ask spread of 0.39% on average. However, as shown in Table 1, call (put) options written on them have a 9.93% (10.38%) quoted bid-ask spread on average. As a result, it is essential that investors take transaction costs into account when they implement option trading signals.

In practice, option investors do not need to pay the full quoted spread to the market makers but the effective spread (ESPR), which is much lower than the quoted spread (QSPR). Zhan et al. (2022) and Muravyev, Pearson, and Van Nieuwerburgh (2020) provide evidence that the actual effective spread is, on average, 55% of the quoted spread, and investors typically pay less than 40% of the conventional effective spread if they can employ execution timing. In this section, we recompute the buy-and-hold delta-hedged option returns corresponding to an assumed effective spread of 25%, 50%, 75%, and 100%. The “No Cost” column in Panel B of Table 2 refers to our main results in Panel A.

Panel B of Table 2 shows that option portfolio strategy based on our ML textual predictor is still profitable for investors even when we assume they need to pay the 100% quoted option spread. Although the (5-1) return spread decreases monotonically with the effective spread, it remains statistically and economically significant for all effective spread values. For example, the return spread is 1.47% (0.87%) per month for call (put) options if we do not consider any transaction cost. When the effective spread increases to 25% of the quoted spread, the return spread decreases to 1.41% (0.80%) for call (put) options. With an effective spread that is 50% (75%) of quoted spread, the return spread further decreases to 1.34% (1.28%) for call options and 0.73% (0.65%) for put options. When we assume that investors must pay the full quoted spread, the return spread remains substantial, with 1.21% (0.58%) for call (put) options.

3.1.2. Double Portfolio Sorts

In this section, we individually control various option or equity characteristics using the double portfolio sorting strategy. We consider 14 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 (ATM) options as in

Goyal and Saretto (2009); (3) stock illiquidity (LOG_AMIHUD), which is the natural logarithm of the monthly average of daily absolute returns divided by daily dollar trading volume following Amihud (2002); (4) option liquidity (OSPREAD) which is the bid-ask option quoted spread scaled by the option mid-price; (5) stock return autocorrelation, calculated as the first-order autocorrelation of underlying stock's return using daily return observations over a past six-month rolling window as in Jeon et al. (2019); (6) Gamma (Gamma) is the sensitivity of an option's delta to the change in the underlying stock price; and (7) Vega (Vega) is the sensitivity of an option's price to its implied volatility.

Following Zhan et al. (2022), we further include the following seven[5] stock characteristics as control variables: (1) the cash-to-assets ratio (CH) proposed by Palazzo (2012), defined as the value of corporate cash holdings over the value of the firm's total assets; (2) analyst earnings forecast dispersion (DISP) proposed by Diether, Malloy, and Scherbina (2002), computed as the standard deviation of annual earnings-per-share forecasts scaled by the absolute value of the average outstanding forecasts; (3) one-year new issues (ISSUE_1Y) proposed by Pontiff and Woodgate (2008), measured as the change in shares outstanding from eleven months ago; (4) the log of the stock price at the end of the current month (LNPRICE) proposed by Blume and Husic (1973); (5) profitability (PROFIT) proposed by Fama and French (2006), calculated as earnings divided by book equity, in which earnings is defined as income before extraordinary items; (6) total external financing (TEF), proposed by Bradshaw, Richardson, and Sloan (2006), is calculated as net share issuance plus net debt issuance minus cash dividends, scaled by total assets; (7) cash flow variance (CFV), as in Haugen and Baker (1996), is computed as the variance of the monthly ratio of cash flow to the market value of equity over the last 60 months. Cash flow is net income plus depreciation and amortization. Panel C of Table 1 reports the summary statistics of the control variables we include here.

Each month, we first sort all options into quintiles based on one of the control variables. Then, we further sort the option sample into quintiles based on the ML textual predictors. Finally, we average returns for each textual predictor quintile across the groups of control variables, yielding five control-variable adjusted quintile returns. Panel A of Table 3 shows that none of the above control variables can subsume the effects of our ML textual predictors. After individually controlling these variables, the return spreads of call options vary from 1.13% to

[5] For sample coverage consideration, we only include seven out of the ten stock characteristics that have predictive power for delta-hedged option returns documented by Zhan et al. (2022). Our main results remain robust when we include all ten stock characteristics.

1.48% per month, and those of put options vary from 0.58% to 0.84% per month. After performing the double portfolio sorting tests, our results remain significant both statistically and economically.

[Insert Table 3]

3.1.3. Fama-Macbeth Regressions

To further affirm the ability of ML predictors based on news media to forecast the cross-section of option returns, we conduct the Fama and MacBeth (1973) regression to test whether the predictive power of textual predictors for delta-hedged option return is statistically significant, especially after simultaneously controlling for existing option return predictors. For each dependent variable (delta-hedged call or put option returns), we run the following cross-sectional regressions where the key independent variable of interest is the SVR textual predicator:

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

where $HPR_{i,t}$ 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 sorts in Section 3.1.2. All independent variables are winsorized at the 0.5th and 99.5th percentiles and standardized cross-sectionally.

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 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 Panel B of Table 3.

The results of Panel B of Table 3 support our claim that ML textual predictors contain useful information about future equity option returns, and their predictability for delta-hedged option returns is robust to various controls. After controlling existing option return predictors, coefficients on ML textual predictors remain both economically and statistically significant. As for the economic magnitude, delta-hedged call (put) option returns increase by 0.259% (0.118%) when the ML textual predictor for call (put) options moves by one standard deviation after controlling existing option predictors. The regression results in Panel B of Table 3 are also consistent with the previous findings of option return predictability in the literature. For example, idiosyncratic volatility is negatively related to delta-hedged option returns, while stock volatility deviation is also a strong predictor to forecast the cross-section of equity option returns.

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 and the benefit of applying the machine learning approach compared to the lexicon-based approach to extract textual information from news media. However, there are potential concerns with the machine learning approach, such as model instability and data mining issues, as multiple hyper-parameters in machine learning approaches can lead to overfitting. To reconcile those concerns, we examine how robust our empirical results are to those hyper-parameters. For our main results based on SVR, there are two primary tuning hyper-parameters: regularization parameter (C) and epsilon (ε). C parameter adds a penalty for each misclassified data point. If C is small, the penalty for misclassified points is low, so a decision boundary with a large margin is chosen at the expense of a greater number of misclassifications. C reflects the strength of the regularization, which can be specified as L_2 penalty. The parameter ε specifies the epsilon-tube within which no penalty is associated in the training loss function for points predicted within a distance ε from the actual value. In our main empirical results, all the hyperparameters are determined through a five-fold cross-validation procedure. We conduct robustness checks by using different parameters of C and ε to train the SVR model. In an untabulated table, we show that the predictive power is robust and significant to different hyperparameters in SVR.

Besides the model parameters, there are also different choices for the input variables. For example, when constructing the word frequency matrix, we need to decide the maximum number of features to input, such as 10,000, 8,000, or 6,000. Also, we can input both unigrams and bigrams into the algorithm. Another input choice is the set of words to be included. While we use all available words in the article in our empirical results, it is interesting to examine which types of words are more important and informative for our target variables. For example, we separate features into different groups based on their part of speech, including nouns, adjectives, adverbs, and verbs. We then run the SVR of delta-hedged option returns on each of those groups and examine which groups are the main drivers of the predictive power of our textual predictors. In the unreported table (available upon request), we find that all types of words are important and necessary to generate significant option return predictability out of the sample. Another choice is the model training period. To investigate whether our results are sensitive to the length of the rolling window to train the model, we change our rolling window from three months to six months, nine months, or twelve months and re-run the SVR. The empirical results are significant and consistent in all cases.

Our main results are based on the support vector regression approach. To verify that our results are not driven by the specific choice of the machine learning approach, 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:

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

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 $x_{i,t-1}$ in the news articles, and λ is a hyper-parameter 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 of time t, we use a rolling window of most recent three months text data to fit the model above to 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:

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

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 elastic net and SVR 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. Different orders of words can indicate totally different meanings in linguistics. To incorporate nonlinearity and interactions among words, we consider more advanced machine learning approaches such as random forest and neural networks. Recent studies Gu, Kelly, and Xiu 2020 show that these methods are helpful in forecasting stock returns.

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:

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

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 the simplest kind of feed-forward network, namely 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 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:

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

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

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

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, 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. To save space, Panel A of Table 4 presents the single portfolio sorting of each textual predictor trained by alternative machine learning approaches. The results of regressions are similar and available upon request.

[Insert Table 4]

Panel A of Table 4 provides consistent evidence that the textual information from news media obtained via alternative machine learning approaches has significant and robust predictive power for delta-hedged equity option returns. Moreover, the trading strategy based on the textual indicators under SVR generates the largest profits among the ML approaches we have implemented. Two things are worth mentioning here: first, the correlations between the alternative ML textual predictors and those from SVR are high, implying that different ML approaches extract similar useful information from news media. For example, textual predictors extracted using neural networks have an average correlation of 0.49 with those from SVR, while predictors extracted from the elastic net and the random forest have correlations of 0.38 and 0.39 with SVR, respectively. Second, the results under elastic net are less robust than those under SVR, random forest, and neural networks. Since the independent variables of word frequencies are very sparse, there are cases when the elastic net could not find a set of non-zero features.

3.2.2. Alternative Constructions of Word Features

In our previous analyses, we use unigram word counts (adjusted by document frequency) to train the machine learning model. The unigram feature is easy to construct with fewer noises, especially in the lexicon-based approach, but also has some limitations. First, the unigram feature does not consider word dependency in different scenarios. A different word adjacent to a certain unigram feature may change its semantic meaning. Second, a model with unigram features lacks interpretability. Many unigram features have meaningful semantics only when they are combined with other words, such as fixed collocations and noun phrases. A possible fix is to extend consecutive words in each feature, such as bigram, trigram, or n-gram. For example, a bigram feature is constructed by including a sequence of two adjacent words within a sentence.

By constructing features in n-grams, we can largely resolve the semantic differences caused by word dependency and model interpretability in the case of unigram features. To check whether our empirical findings are special to the choice of unigram feature, we re-train our SVR model to forecast equity option returns using various n-gram features described above. We consider three types of n-grams (bigram, trigram, and fivegram) and train our model. When processing the text data into n-gram features, we first divide each newspaper article to the sentence level and make necessary adjustments, such as excluding stop words, reducing a word to its word stem, and switching a word to its base root mode (i.e., lemmatization). We then treat each n-gram as a new feature and count its feature frequency within the whole article. Following our steps in Section 2, we adjust the count numbers by the process of *tf–idf* (term frequency-inverse document frequency). 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 (6). The empirical results are provided in Panel B of Table 4. To save space, Panel B of Table 4 presents the single portfolio sorting test of each textual predictor trained by alternative machine learning approaches.

Panel B of Table 4 provides consistent evidence that textual predictors extracted from news media using different n-gram features are still significant predictors for delta-hedged equity option returns. It is worth noting that although a larger n for the n-gram feature (e.g., fivegram) can provide a more interpretable model, the feature might become noisy and less combinable. More specifically, in our case of fivegram, there will be fewer identical features that will have the same five words in the same order. Consequently, the algorithm may classify two semantic fivegram features as different, even if they differ by only one word. To solve this issue, we apply some recently developed word embedding algorithms to measure the similarity between two different fivegram features and classify similar fivegrams as the same feature. Word embedding algorithms have been demonstrated to help identify semantic meanings in finance (see, e.g., Jha, Liu, and Manela 2020).

To combine similar features together for our case of fivegram, we first use some well-trained word embedding models, such as Word2Vec, GloVe, and Bidirectional Encoder Representations from Transformers (BERT), to transfer each word to a vector with a fixed dimension (e.g., 100, 300, 1000, etc.). Most of the word embedding models are pre-trained neural network models that project words to a high-dimensional vector space so that words with relevant semantic meanings can be closer to each other in the vector space. Word embedding also allows us to construct a vector representation for a certain n-gram feature and measure its semantic similarity with other n-gram features in the projected vector space. More specifically, we transfer each word j to a 300×1 vector w_t^j using Word2Vec. For each fivegram feature g, its word embedding s_t^g can be proxied by the average of the word embedding vectors for the words belonging to the fivegram:

$$s_t^g \equiv \left[s_t^{g1}, \dots, s_t^{g300}\right] = \frac{1}{5}\sum_{j \in g} w_t^j. \tag{16}$$

The technique is well-known as the continuous bag of words (CBOW) approach. Although CBOW ignores word orders within the fivegram, the method is demonstrated to be an effective embedding model in practice. After obtaining the average word embedding vector for each fivegram, we compute the cosine similarity between s_t^g and the word embedding vector of any other fivegrams in the training dataset at time t:

$$c_t^{gh} = \frac{s_t^g \cdot s_t^h}{\left|s_t^g\right|\left|s_t^h\right|} = \frac{\sum_d s_t^{gd} s_t^{hd}}{\sqrt{\sum_d \left(s_t^{gd}\right)^2}\sqrt{\sum_d \left(s_t^{hd}\right)^2}}, \tag{17}$$

where $d = 1, \dots, 300$ stands for the elements in the vector of s_t^g and s_t^h. The cosine similarity helps us find correlated fivegram features (in the sense of semantics) and combine them into one feature, thus reducing estimation noises for large n-gram features. In our case, we combine two fivegrams into one feature if the cosine similarity of their word embedding vectors is greater than 0.9. The rest of the empirical work proceeds in the same way as in the case of the unigram feature previously.

3.2.3. Alternative Option Samples and Return Construction

In this section, we investigate the robustness of our ML textual predictors in different option samples and the alternative construction of option returns. In our main analysis, we consider options whose moneyness is closest to 1 within the range from 0.8 to 1.2. Here, we define in-the-money (ITM) call options as options whose moneyness is closest to 0.9 within the range from 0.8 to 1.0, and out-of-money (OTM) call options as options whose moneyness is closest to 1.1 within the range from 1.0 to 1.2. ITM puts and OTM puts are defined correspondingly, i.e., ITM puts have moneyness closest to 1.1 within the range from 1.0 to 1.2 while OTM puts have moneyness closest to 0.9 within the range from 0.8 to 1.2.

Apart from different moneyness, we also consider different maturities. In our baseline results, at the end of month *t*, our sample includes options expiring in month *t*+2, meaning they have 1.5 months to expire. Here we include at-the-money (ATM) options expiring in *t*+3 and *t*+4, respectively. This sample is equivalent to a sample that includes options with 2.5 months and 3.5 months to expire, respectively.

Our ML textual predictors are trained by inputting buy-and-hold delta-hedged option returns. We investigate whether they still have predictive power on daily-rebalanced delta-hedged option returns. The daily-rebalanced delta-hedged option returns are defined as the total dollar gains of daily-rebalanced option positions scaled by the absolute values of the initial costs. For example, the dollar gain of a daily-rebalanced call option position is defined as:

$$\Pi_{t,t+\tau} = C_{t+\tau} - C_t - \sum_{n=0}^{N-1} \Delta_{c,t_n}\left(S_{t_{n+1}} - S_{t_n}\right) - \sum_{n=0}^{N-1} \frac{a_n r_{t_n}}{365}\left(C_{t_n} - \Delta_{c,t_n} S_{t_n}\right), \tag{18}$$

where Δ_{c,t_n} is the call delta of the call option on the date t_n, $r_{t_n}$is the annualized risk-free rate on the date 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. With a zero-net investment initial position, the delta-hedged option gain $\Pi_{t,t+\tau}$ is the excess dollar return of the delta-hedged option. To make option returns comparable across stocks, we scale the dollar return by the initial costs of the portfolio, i.e., $\Delta_{c,t}S_t - C_t$ for call options and $P_t - \Delta_{p,t}S_t$ for puts.

[Insert Table 5]

The predictability of our ML textual predictors remains significant in alternative option samples with different moneyness and maturity. As shown in Table 5, the (5-1) return spreads are statistically and economically significant, ranging from 0.74% to 1.78% for call options and 0.35% to 1.05% for put options. Our ML textual predictors can also predict daily rebalanced delta-hedged option returns with a return spread of 0.69% (0.39%) for call (put) options, suggesting that our results are not driven by the underlying stocks.

4. Interpretations of Textual Predictors and Economic Mechanisms

4.1. Nature of the Textual Information

4.1.1. Important Words in Constructing Textual Predictors

We have provided comprehensive evidence that qualitative information from news media is useful for predicting delta-hedged option returns. Yet, the nature of such information extracted using machine learning approaches is ex-ante unclear. In this section, we aim to shed some lights on the interpretation of the SVR textual predictors. Following Manela and Moreira (2017), we measure the importance of each word feature by the magnitude of its coefficient in SVR. We select the top 1,000 words with the largest magnitude of the coefficients as the important information set. Unlike a pre-specified dictionary used to quantify sentiment or political risk in the literature, this dictionary of important words under SVR is time varying. We define the overall feature importance for each word as the fraction of time (percentage of months) in which the word appears in the top 1,000 most important words.

We classify each word into various topics using pre-specified dictionaries, such as the LM dictionary, the GI dictionary, the political dictionary, and the macroeconomic dictionary. The politic and macroeconomics news are the main topics in addition to firm-specific news in newspaper articles about companies. The sentiment dictionary is the combination of the GI dictionary and the LM dictionary, the political dictionary is from Hassan et al. (2019) and constructed by us, and the macroeconomic dictionary is based on Bloom (2014) and Bybee, Kelly, Manela, and Xiu (2020). For the remaining words selected by SVR as important features, we refer to them as firm-idiosyncratic words reflecting firm-specific news media information. For example, some words are only related to certain industries, product markets, or certain companies. Table A1 provides a sample list of words belonging to each topic. For concision and simplicity, we list the top 20 words for each important feature set.

We also examine the feature importance for the elastic net, random forest, and neural networks. The definition of feature importance for the elastic net is the same as that for SVR, while the feature importance for the random forest is obtained by computing the proportion of each word selected among all the random decision trees generated by the model. For the feature importance for neural networks, because of hidden layers and activation functions, we cannot use coefficients directly as the measure of its feature importance. Instead, we apply an algorithm to approximate the feature importance, namely feature permutation importance. Feature permutation importance is a model-agnostic global explanation method that provides insights into a machine learning model's behavior. It estimates and ranks feature importance based on each feature's impact on the trained machine learning model's predictions. The algorithm can be applied to any black-box estimator by evaluating how the prediction error increases when a feature is unavailable. In practice, instead of removing features and retraining the estimator for each feature, the algorithm randomly shuffles the feature values, effectively adding noise to the feature. Then the prediction error of the new dataset is compared with the prediction error of the original dataset. If the model heavily relies on those features being shuffled to better forecast the target variable, those features are identified as important inputs. Otherwise, the features are not important.

Table 6 shows the average feature importance of each topic. To count the proportion of important words from each topic, we first select the top 1,000 words with the largest coefficient magnitudes from the SVR model in each month. Then we count the percentage of words in the feature important sets that belong to each topic and average it over time. The results are provided in Table 6.

[Insert Table 6]

As can be seen from Table 6, the proportion of words for each topic shows that most of the textual information extracted from news media is about sentiment words from the GI dictionary and the LM dictionary, which accounts for more than 60% of the selected important features. The rest of the important contribution is from firm-idiosyncratic words accounting for around 20%. On the contrary, macroeconomic information, such as economic policies and politics accounts for a small proportion of the information sets that explain the cross-sectional variations of delta-hedged equity option returns. The results for n-gram features are similar to those reported in Table 6 for unigram features: in terms of the relative explanatory power of the n-gram features, most of the important contributions come from sentiment words and firm-specific words rather than macroeconomic information.

While sentiment words account for a large proportion of feature importance, firm-idiosyncratic words also play an essential role. In Panel A of Table A2, we further show that most of those firm-idiosyncratic words are accompanied by words from the GI and the LM dictionaries. Specifically, around 75% of the selected firm-idiosyncratic words appear in a sentence that includes a word from either the GI or the LM dictionary.

To obtain the results reported in Panel A of Table A2, we construct a synthetic bigram using both the firm-idiosyncratic words and the sentiment words selected from SVR to visualize the semantic interpretation of our model. The bigram is combined at the sentence level if and only if it contains both a firm-idiosyncratic word and a sentiment word. To combine the two words, we keep the firm-idiosyncratic word always as the first, followed by the sentiment word, so that the bigram is more interpretable. For example, in the sentence, "We are confident our accounting treatment was correct.", *treatment* and *confident* are selected from the groups of firm-idiosyncratic words and sentiment words respectively. The combination is recorded as *treatment-confident*. In Panel A of Table A2, we list those top bigrams extracted by SVR that consist of both firm-idiosyncratic words and sentiment words. Most bigrams are related to either positive or negative firm-specific information or disagreement/uncertainty about firms' future performance. For example, among these selected combinations, many bigrams involve modal verbs, such as *would*, *could*, *may*, or *might*, indicating some uncertainties or disagreements.

As another attempt to interpret our textual predictors, we train the SVR model using fivegram features (which are more interpretable than unigram features) described in Section 3.2.2. We list in Panel B of Table A2 those top fivegrams ranked by the feature importance rule for SVR. Consistent with the observation in Panel A of Table A2, most of the fivegrams include words in the GI dictionary and LM sentiment dictionary. Moreover, many fivegrams involve firm-specific words, such as chief executive officer, earnings, and analyst.

In an untabulated table, we find that although the delta-hedged call and put options are highly correlated, the textual predictors extracted based on call and put options are quite different. For example, among the top 1,000 words of feature importance, only around 20% of words overlap between call and put option textual predictors. Low overlap between call and put option dictionaries could indicate that the textual information we extracted from news media to forecast call and put option returns are different. However, sentiment-related and firm-idiosyncratic words account for the largest portion of all important word features for both call and put options.

4.1.2. Relationships between SVR Predictors and Sentiment Measures

As previously documented, most of the words that importantly contribute to the ML textual predictors are sentiment related, which motivates us to examine whether those ML textual predictors are related to investor sentiment measures. We apply the lexicon-based approach to construct investor sentiment measures using some pre-specified dictionary, such as GI dictionary in Tetlock (2007) and Loughran-McDonald (LM) finance-specific dictionary in Loughran and McDonald (2011). Based on the Harvard-IV-4 psychosocial dictionary and the Lasswell value dictionary, the GI dictionary reflects Charles Osgood's semantic differential findings regarding basic language universals. Recent studies show that the GI dictionary and the LM dictionary are both good approximations of sentiment measures for finance Shapiro, Sudhof, and Wilson 2022 but the LM dictionary is more effective in computing sentiment information of finance-related documents such as SEC EDGAR files and IPO prospectuses Jegadeesh and Wu 2013. We combine the GI dictionary and the LM dictionary since the news media we analyze covers broader topics than finance. Nevertheless, our results are robust if we use only one of the GI and the LM dictionaries.

We construct two sentiment-related measures: sentiment level and sentiment disagreement. First, following previous studies, we compute the sentiment score that reflects the tone of each news article using the difference between the number of positive and negative words, scaled by the total number of sentiment words in each article. Then, during each time period t (e.g., a given month) and for each firm i, the media sentiment on the firm is measured as the average sentiment score of all articles on firm i within this time period:

$$GILM_{i,t}^{1} = \frac{1}{H_{i,t}} \sum_{h=1}^{H_{i,t}} \frac{\#(Positive)_{i,h} - \#(Negative)_{i,h}}{\#(Positive)_{i,h} + \#(Negative)_{i,h}}, \ \ i = 1, \ldots, N_t, \tag{19}$$

where $\#(Positive)$ or $\#(Negative)$ denotes the number of positive or negative words detected based on the GI and LM dictionaries in each article. $H_{i,t}$ stands for the total number of articles covering firm i in time period t. A higher $GILM^1$ indicates a more positive view in the news media about the firm and vice versa.

Second, to capture the range of different sentiment among news articles on each firm during a time period t, we follow Cookson and Niessner (2019) and define sentiment disagreement as the standard deviation of article-level sentiment:

$$GILM_{i,t}^{2} = \sqrt{\frac{1}{H_{i,t}} \sum_{h=1}^{H_{i,t}} \left(\frac{\#(Positive)_{i,h} - \#(Negative)_{i,h}}{\#(Positive)_{i,h} + \#(Negative)_{i,h}} - GILM_{i,t}^{1} \right)^2}, \ \ i = 1, \ldots, N_t. \tag{20}$$

As a special case, when there is only one article for the firm at a given point of time, $GILM_{i,t}^{2}$

is measured as the absolute value of $GILM_{i,t}^{1}$. A higher $GILM_{i,t}^{2}$ indicates more divergent sentiment about the company from the news media.

Alternatively, we construct two similar sentiment related measures using the Composite Sentiment Score provided by a commercial database RavenPack News Analytics, a leading media database widely used in the finance literature. For each company and in each month, we define Raven^{1} (resp. Raven^{2}) as the mean (resp. standard deviation) of the Composite Sentiment Score of all news articles covering the company during the given month. We only retain news articles with a relevance score equaling 100.

In Table 7, we regress the lexicon-based or RavenPack sentiment measures averaged over the most recent three months (so as to match the training window of machine-learning algorithms) on SVR predictors. Panel A of Table 7 shows that SVR predictors are strongly associated with all sentiment-related measures at the 1% significance level. Specifically, SVR predictors are positively related to sentiment level measures and negatively related to sentiment disagreement measures.

[Insert Table 7]

Although SVR predictors are significantly correlated with sentiment-related measures, they capture more information than measures based on dictionaries (GILM^{1}, GILM^{2}) or RavenPack (Raven^{1}, Raven^{2}) in predicting cross-section of option returns. In Panel B of Table 7, we compare the (5-1) portfolio returns generated by SVR predictors, lexicon-based sentiment measures, and RavenPack sentiment measures. The absolute values of (5-1) portfolio returns based on SVR predictors are much larger than those based on traditional sentiment-related measures. For example, the largest absolute return spread generated by traditional sentiment measures is 0.68% for call options and 0.59% for put options, much less than those generated by SVR predictors (1.47% for call options and 0.87% for put options). Further in Panel C of Table 7, we regress the delta-hedged option returns on SVR predictors when controlling each sentiment-related measure. We find that the predictability of SVR predictors cannot be subsumed by any sentiment-related measure, indicating that SVR predictors capture information beyond the scope of sentiment-related information. Later in Section 4.3, we show that SVR predictors are related to the uncertainty about the implied volatility and can significantly predict future implied volatility changes.

Our findings indicate that when extracting information from unstructured data to predict option returns, our supervised machine-learning approaches perform better than traditional methods based on dictionaries. Signals obtained by machine-learning approaches are data-driven and capture information from various perspectives, while traditional dictionaries-based methods focus on a specific factor that might affect option pricing (e.g., sentiment).

To further illustrate the important information contained in non-sentiment words for predicting option returns, we form an alternative SVR predictor (SVR_GILM) based on news articles but keeping only words from the GI dictionary and the LM dictionary. The predictive performance of call and put option returns under this alternative approach is shown in column (6) of Table 7, Panel B. The return spread generated by SVR_GILM is only 0.12% (0.18%) per month for call (put) options. The predictive power of SVR predictors almost vanishes when only sentiment-related words are included, indicating the importance of non-sentiment related words.

4.2. Potential Economic Channels

In this section, we investigate potential economic channels that could explain the predictability of SVR predictors. Specifically, we consider the possibilities that machine learning approaches extract useful information from news media that capture uncertainty about firm volatility, jump risk, or option demand pressure which have been shown to be significantly related to expected option returns. We test whether SVR predictor is still significantly to delta-hedged option returns after controlling for proxies of uncertainty about firm volatility, jump risk, option demand pressure.

First, delta-hedged options are sensitive to the underlying stock volatility, and the uncertainty about it expose option market makers to higher risk. Recent studies by Huang, Schlag, Shaliastovich, and Thimme (2019) and Cao, Vasquez, Xiao, and Zhan (2022) show that the volatility of volatility (VOL-of-VOL) is an important determinant of delta-hedged option return both theoretically and empirically. Following Cao et al. (2022), we measure the volatility of volatility as the standard deviation of the percentage changes of the daily implied volatility (VOIV) or realized volatility (VORV) of a given month. The implied volatility is obtained from the Volatility Surface file of OptionMetrics. The implied volatility of a given stock is the average of its call and put option’s implied volatilities with absolute delta of 0.5 and expiration of 30 days. The realized volatility is the five-minute intraday volatility calculated using the TAQ data. We record prices every five minutes starting from 9:30 EST and construct five-minute log-returns. We use the last recorded price within each five-minute period to calculate the log return.

Second, although delta-hedged option portfolios are immune from small price changes of the underlying stocks, large price movements can cause the delta-hedging strategy to be ineffective. Green and Figlewski (1999) argue that option writers charge a premium for the jump risk, and Broadie, Chernov, and Johannes (2009) show that considering a jump risk premium can better understand the index option returns. Following Bakshi and Kapadia (2003) and Bakshi, Kapadia, and Madan (2003), we use the model-free implied skewness (MFIS) and kurtosis (MFIK) to measure the jump risk.

Third, as shown by Bollen and Whaley (2004) and Gârleanu, Pedersen, and Poteshman (2009), option market makers cannot perfectly hedge their inventories and are thus exposed to inventory risk. Higher demand pressure from end-users causes option market makers to face higher inventory risk, and thus market makers charge a premium for the inventory risk. Muravyev (2016) empirically shows the importance of inventory risk in option pricing. We use two measures to proxy for the option demand pressure. First, we use the total market value of all options on a stock (open interest times option price), scaled by the underlying stock's market value to proxy for the option demand pressure. Second, we use the order imbalance of options to measure the demand pressure from end-users. Utilizing the Open/Close data from The Chicago Board Options Exchange (CBOE) and the International Securities Exchange (ISE), we construct the option order imbalance using the following equation:

$$OIB_{i,t} = (Open\,Buy_{i,t} - Open\,Sell_{i,t})/\left(Open\,Buy_{i,t} + Open\,Sell_{i,t}\right), \tag{21}$$

where Open Buy (Sell) is the total volume of buyer (seller)-initiated orders to open new positions. For firm *i* and month *t*, the $Open\,Buy_{i,t}$ $(Open\,Sell_{i,t})$ is the total buyer-(seller) initiated volume of all tradable options written on firm *i* across month *t*. Option demand pressure measures are separately calculated for call options and put options.

In Panel A (B) of Table 8, we individually control for each proxy in the Fama-Macbeth regression for call (put) options. Although each proxy has a significant predictive power for delta-hedged option returns, none of them can subsume the significance of the SVR predictor. The coefficient of the SVR predictor barely changes when we control proxies for jump risk or option demand pressure. The maximum reduction of SVR predictor' coefficient occurs after we control for the volatility of implied volatility, and the reduction is 16% (22%) for call (put) options.[6] These results indicate that the underlying economic mechanism of SVR predictors is not related to jump risk or option demand pressure from end-users. In comparison, uncertainty about the volatility accounts for some of the option return predictability by the SVR predictor. In the next section, we further investigate the information content of the SVR predictor. In particular, we show that SVR predictor contains valuable information about future volatility changes.

[Insert Table 8]

4.3. Information Contents of Textual Predictors

To identify the information contents of textual predictors, we further examine whether they can predict implied volatility change or underlying stock fundamentals. Our motivation starts from the following decomposition of the buy-and-hold delta-hedged call option returns:

$$HRP_{i,t+1}^{C} = \frac{(C_{i,t+1} - C_{i,t}) - \Delta_{i,t}^{\mathrm{C}}(S_{i,t+1} - S_{i,t})}{\left|C_{i,t} - \Delta_{i,t}^{\mathrm{C}} \times S_{i,t}\right|} = w \times R_{option} - (1 + w) \times R_{stock}, \qquad (22)$$

where $w = C_{i,t}/\left|C_{i,t} - \Delta_{i,t}^{\mathrm{C}} \times S_{i,t}\right|$. The first part is related to the raw return of call options, while the second is related to the return of the underlying stocks. Therefore, the buy-and-hold delta-hedged call option return is the weighted average of the returns for the call option and the underlying stock. A similar decomposition applies to buy-and-hold delta-hedged put option returns. Since option price is an increasing function of implied volatility, option returns is positively related to the implied volatility change over the holding period. Moreover, the predictability of textual predictors can also stem from its information content about the underlying firm fundamentals and stock returns.

Thus, we empirically test the relations between SVR predicators and future implied volatility change or the underlying stock return and future earnings surprise. First, we investigate whether SVR predictor can predict the implied volatility change over the following month. For each stock in our sample, we obtain the implied volatility for a call and a put with absolute delta of 0.5 and expiration of 30 days from the OptionMetrics volatility surface database. The implied volatility change is calculated as the percentage difference between the

[6] When including option OIB as the control variable, the without-control coefficient of SVR predictors change to 0.349 (0.242) for call (put) options due to sample difference. Therefore, the reduction is only 1.7% (0.4%) for call (put) options.

next month's and the current month's implied volatility.

Second, we test whether SVR predictors contain information that helps predict future earnings surprises or stock excess returns. Following Livnat and Mendenhall (2006), we calculate the standardized unexpected earnings (SUE) based on the assumption that earnings per share (EPS) follows a seasonal random walk, where the best expectation of the EPS in quarter t is the firm's reported EPS in the same quarter of the previous fiscal year. Hence, the SUE is given by the following equation:

$$SUE_{i,t} = \frac{\left(X_{i,t} - X_{i,t-4}\right)}{P_{i,t}}, \quad i = 1, \dots, N_t, \tag{23}$$

where $X_{i,t}$ is primary earnings per share (EPS) before extraordinary items for firm i in quarter t, and $P_{i,t}$ is the price per share for firm i at the end of quarter t from Compustat. $X_{i,t}$ and $P_{i,t}$ are unadjusted for stock splits, but $X_{i,t-4}$ is adjusted for any stock splits and dividends during the period $[t-4, t]$ using the Compustat adjustment factor (AJEXQ). "Special items" in Compustat are excluded from the calculation of the SUE. The higher SUE is, the more positive the earnings surprise is. Stock excess return is defined as the difference between the raw stock return and the risk-free return over the next month.

To check the information contents of SVR predictors, we run predictive regressions with the implied volatility change or stock excess return as dependent variables $Z_{i,t}$:

$$Z_{i,t} = a + b \times SVR_{i,t-1} + \sum_{j=1}^{M} \gamma_t^j X_{i,t-1}^j + \epsilon_{i,t}, \quad i = 1, \dots, N_t, \tag{24}$$

where $SVR_{i,t-1}$ is the SVR predictor (i.e., $\hat{r}_{i,t}$) for firm i at time $t-1$, and $X_{i,t-1}^j$ are control variables. We run univariate regression of SUE or stock excess return on SVR predictors. When regressing next month implied volatility changes on SVR predictors, we control the implied volatility change in the current month, i.e., the percentage difference between the current month's and the previous month's implied volatility.

[Insert Table 9]

Table 9 shows that the SVR predictors are not significantly related to future earnings surprises or stock excess returns. Instead, SVR predictors show significant predictability to future implied volatility changes. This suggests that textual information from new media that predict delta-hedged option returns do not come from the underlying fundamentals.

On the other hand, we find that after controlling the lagged implied volatility change, SVR predictors still significantly predict future implied volatility change at 1% significance level for both call options and put options. This combined with the earlier result in Section 4.2 (uncertainty about the implied volatility has the largest explanatory power for the option return predictability of SVR predictors) suggest that SVR predictors' ability to predict delta-hedged option returns is intimately related to its information content about future volatility. Our work suggests that valuable and useful insights about stock volatility can be gleaned from news media coverage of the company using machine learning algorithms.

We acknowledge that it is a challenging task to interpret signals obtained by machine-learning algorithms and there could be additional channels underlying the option returns predictability by textual predictors we extract from news media using machine learning algorithms. We leave further explorations to future work.

5. Conclusion

In this paper, we study whether and how textual information from news media could predict the cross-section of delta-hedged option returns. First, we find that the textual information extracted from news media using machine learning techniques has a strong predictive power for delta-hedged option returns. The results are robust to different methodologies and controlling for various option return predictors documented in the literature. Second, a large proportion of the predictive power of the textual predictor is qualitative sentiment-related information, as the important word features from trained ML models have the largest overlap with sentiment dictionaries. Third, the predictive power of textual predictors for delta-hedged option return is unrelated to future returns of the underlying stock or future company earnings. Instead, we find that news media contains useful information about future change in stock volatility which helps to explain the predictive power of textual predictors for delta-hedged option return.

Overall, our results demonstrate that machine-learning methods can extract useful information from news media that are both statistically and economically significant predictors of option returns. Our paper provides a novel angle to predict option returns and illustrate the importance of incorporating information from unstructured data when pricing options. Future research could explore more advanced machine learning approaches (such as recurrent neural network and convolutional neural network) and incorporate dependency across words in a document to extract information from text data. More work is needed to better understand the underlying mechanisms for the delta-hedged option return predictability by textual information from news media. This exercise would also shed new lights on the cross-sectional determinants of expected delta-hedged option returns.

References

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

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang. 2006. The Cross-Section of Volatility and Expected Returns. *The Journal of Finance* 61: 259-99.

Bakshi, G., and N. Kapadia. 2003. Delta-Hedged Gains and the Negative Market Volatility Risk Premium. *Review of Financial Studies* 16: 527-66.

Bakshi, G., N. Kapadia, and D. Madan. 2003. Stock Return Characteristics, Skew Laws, and the Differential Pricing of Individual Equity Options. *Review of Financial Studies* 16: 101-43.

Bali, T. G., H. Beckmeyer, M. Moerke, and F. Weigert. 2021. Option Return Predictability with Machine Learning and Big Data. *SSRN Electronic Journal*.

Bloom, N. 2014. Fluctuations in Uncertainty. *Journal of Economic Perspectives* 28: 153-76.

Blume, M. E., and F. Husic. 1973. Price, Beta, and Exchange Listing. *The Journal of Finance* 28: 283-99.

Bollen, N. P. B., and R. E. Whaley. 2004. Does Net Buying Pressure Affect the Shape of Implied Volatility Functions? *The Journal of Finance* 59: 711-53.

Bradshaw, M. T., S. A. Richardson, and R. G. Sloan. 2006. The relation between corporate financing activities, analysts' forecasts and stock returns. *Journal of Accounting and Economics* 42: 53-85.

Broadie, M., M. Chernov, and M. Johannes. 2009. Understanding Index Option Returns. *Review of Financial Studies* 22: 4493-529.

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu. 2020. The Structure of Economic News. *National Bureau of Economic Research Working Paper Series* No. 26648.

Cao, J., and B. Han. 2013. Cross section of option returns and idiosyncratic stock volatility. *Journal of Financial Economics* 108: 231-49.

Cao, J., A. Vasquez, X. Xiao, and X. Zhan. 2022. Why Does Volatility Uncertainty Predict Equity Option Returns? *SSRN Electronic Journal*.

Chinco, A., A. D. Clark-Joseph, and M. A. O. Ye. 2019. Sparse Signals in the Cross-Section of Returns. *The Journal of Finance* 74: 449-92.

Cookson, J. A., and M. Niessner. 2019. Why Don't We Agree? Evidence from a Social Network of Investors. *The Journal of Finance* 75: 173-228.

Cortes, C., and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20: 273-97.

Coval, J. D., and T. 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: 303-14.

Diether, K. B., C. J. Malloy, and A. Scherbina. 2002. Differences of Opinion and the Cross Section of Stock Returns. *The Journal of Finance* 57: 2113-41.

Dong, X. I., Y. A. N. Li, D. E. Rapach, and G. Zhou. 2021. Anomalies and the Expected Market Return. *The Journal of Finance* 77: 639-81.

Eisdorfer, A., A. Goyal, A. Zhdanov, and A. Boulatov. 2022. Limited Attention and Option Prices. *SSRN Electronic Journal*.

Engle, R. F., S. Giglio, B. Kelly, H. Lee, and J. Stroebel. 2020. Hedging Climate Change News. *The Review of Financial Studies* 33: 1184-216.

Fama, E. F., and K. R. French. 2006. Profitability, investment and average returns. *Journal of Financial Economics* 82: 491-518.

———. 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116: 1-22.

Fama, E. F., and J. D. MacBeth. 1973. Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy* 81: 607-36.

Frankel, R., J. Jennings, and J. Lee. 2022. Disclosure Sentiment: Machine Learning vs. Dictionary Methods. *Management Science* 68: 5514-32.

Gârleanu, N., L. H. Pedersen, and A. M. Poteshman. 2009. Demand-Based Option Pricing. *Review of Financial Studies* 22: 4259-99.

Goyal, A., and A. Saretto. 2009. Cross-section of option returns and volatility☆. *Journal of Financial Economics* 94: 310-26.

Green, T. C., and S. Figlewski. 1999. Market Risk and Model Risk for a Financial Institution Writing Options. *The Journal of Finance* 54: 1465-99.

Gu, S. H., B. Kelly, and D. C. Xiu. 2020. Empirical Asset Pricing via Machine Learning. *Review of Financial Studies* 33: 2223-73.

Hassan, T. A., S. Hollander, L. van Lent, and A. Tahoun. 2019. Firm-Level Political Risk: Measurement and Effects*. *The Quarterly Journal of Economics* 134: 2135-202.

Haugen, R. A., and N. L. Baker. 1996. Commonality in the determinants of expected stock returns. *Journal of Financial Economics* 41: 401-39.

Hornik, K., M. Stinchcombe, and H. White. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks* 2: 359-66.

Huang, D., C. Schlag, I. Shaliastovich, and J. Thimme. 2019. Volatility-of-Volatility Risk. *Journal of Financial and Quantitative Analysis* 54: 2423-52.

Jegadeesh, N., and D. Wu. 2013. Word power: A new approach for content analysis. *Journal of Financial Economics* 110: 712-29.

Jeon, Y., R. Kan, and G. Li. 2019. Stock Return Autocorrelations and the Cross Section of Option Returns. *SSRN Electronic Journal*.

Jha, M., H. Liu, and A. Manela. 2020. Does Finance Benefit Society? A Language Embedding Approach. *SSRN Electronic Journal*.

Ke, Z. T., B. T. Kelly, and D. Xiu. 2019. Predicting Returns With Text Data. *National Bureau of Economic Research Working Paper Series* No. 26186.

Kelly, B., A. Manela, and A. Moreira. 2021. Text Selection. *Journal of Business & Economic Statistics* 39: 859-79.

Livnat, J., and R. R. Mendenhall. 2006. Comparing the Post-Earnings Announcement Drift for Surprises Calculated from Analyst and Time Series Forecasts. *Journal of Accounting Research* 44: 177-205.

Loughran, T. I. M., and B. McDonald. 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance* 66: 35-65.

Manela, A., and A. Moreira. 2017. News implied volatility and disaster concerns. *Journal of Financial Economics* 123: 137-62.

Muravyev, D. 2016. Order Flow and Expected Option Returns. *The Journal of Finance* 71: 673-708.

Muravyev, D., N. D. Pearson, and S. Van Nieuwerburgh. 2020. Options Trading Costs Are Lower than You Think. *The Review of Financial Studies* 33: 4973-5014.

Newey, W. K., and K. D. West. 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55.

Palazzo, B. 2012. Cash holdings, risk, and expected returns. *Journal of Financial Economics* 104: 162-85.

Pontiff, J., and A. Woodgate. 2008. Share Issuance and Cross-sectional Returns. *The Journal of Finance* 63: 921-45.

Ramachandran, L. S., and J. Tayal. 2021. Mispricing, short-sale constraints, and the cross-section of option returns. *Journal of Financial Economics* 141: 297-321.

Schölkopf, B., and A. J. Smola. 2002. *Learning with kernels support vector machines, regularization, optimization, and beyond*. Cambridge, Mass: MIT Press.

Shapiro, A. H., M. Sudhof, and D. J. Wilson. 2022. Measuring news sentiment. *Journal of*

Econometrics 228: 221-43.

Tetlock, P. C. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance* 62: 1139-68.

———. 2010. Does Public Financial News Resolve Asymmetric Information? *Review of Financial Studies* 23: 3520-57.

Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy. 2008. More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance* 63: 1437-67.

Tian, M., and L. Wu. 2021. Limits of Arbitrage and Primary Risk Taking in Derivative Securities. *SSRN Electronic Journal*.

Zhan, X., B. Han, J. Cao, and Q. Tong. 2022. Option Return Predictability. *The Review of Financial Studies* 35: 1394-442.

Table 1 Summary Statistics

This table reports the descriptive statistics of option returns, option characteristics, textual predictors, and equity characteristics. The sample period is from February 1996 to December 2018. Panel A (B) reports the pooled summary of delta-hedged call (put) option returns and the characteristics of call (put) options involved. 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. The position is held for 1 month or until option maturity. 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. Moneyness is the ratio of option strike price to stock price. Days to maturity is the number of calendar days until the option expires. Gamma is the Black-Scholes option gamma. Vega is the Black-Scholes option vega. Option bid-ask spread is the ratio of the difference between ask and bid quotes of option to the midpoint of the bid and ask quotes at the end of each month. Panel C reports the time-series average of cross-sectional statistics of textual predictors and equity characteristics (all independent variables are winsorized each month at the 1% level). Call_SVR (Put_SVR) is the textual predictor extracted from news media using support vector regression model. AUTO is the first-order autocorrelation of underlying stock's return as in Jeon et al. (2019). CASH is the cash-to-assets ratio as in Palazzo (2012). CFV is the cash flow variance as in Haugen and Baker (1996). DISP is the analyst earnings forecast dispersion, as in Diether et al. (2002). ISSUE_1Y represents 1-year new issues as in Pontiff and Woodgate (2008). IVOL is the idiosyncratic volatility computed as in Ang et al. (2006). LNPRICE is the log of the underlying stock price at the end of last month. LOG_AMIHUD is the natural logarithm of the illiquidity measure from Amihud (2002). PROFIT is the profitability as in Fama and French (2006). TEF is total external finance. HV-IV is the difference between realized volatility and implied volatility as in Goyal and Saretto (2009). Panel D reports the cross-sectional Pearson correlations of textual predictors and various characteristics of options and stocks. All variables are winsorized each month at the 1% level. We report the cross-sectional correlations each month and report the time-series average of these correlations.

Panel A: Pooled Summary of Delta-hedged Call Option Returns and Option Characteristics (50,888 observations)

	Mean	Standard deviation	10th percentile	Lower quartile	Median	Upper quartile	90th percentile
Buy & hold until month-end (%)	-2.67	4.76	-7.22	-4.56	-2.61	-0.92	1.49
Buy & hold until maturity (%)	-8.47	6.24	-15.91	-10.77	-7.09	-4.54	-2.76
Moneyness (%)	99.98	3.20	96.34	98.36	100.00	101.59	103.65
Days to Maturity	50	2	47	50	50	51	52
Gamma	0.09	0.05	0.03	0.05	0.08	0.11	0.16
Vega	0.14	0.01	0.13	0.14	0.14	0.15	0.15
Quoted option bid-ask spread (%)	9.93	8.82	2.39	4.44	7.79	12.66	19.05

Panel B: Pooled Summary of Delta-hedged Put Option Returns and Option Characteristics (50,888 observations)

	Mean	Standard deviation	10th percentile	Lower quartile	Median	Upper quartile	90th percentile
Buy & hold until month-end (%)	-2.03	3.97	-5.99	-3.81	-2.11	-0.54	1.71
Buy & hold until maturity (%)	-7.75	5.01	-14.14	-9.93	-6.74	-4.47	-2.81
Moneyness (%)	99.96	3.22	96.28	98.34	100.00	101.57	103.63
Days to Maturity	50	2	47	50	50	51	52
Gamma	0.09	0.05	0.03	0.05	0.08	0.11	0.16
Vega	0.14	0.01	0.13	0.14	0.14	0.15	0.15
Quoted option bid-ask spread (%)	10.38	9.45	2.41	4.65	8.00	13.33	20.00

Panel C: Equity Characteristics (Time-series Average of Cross-sectional Statistics)

	Mean	Standard deviation	10th percentile	Lower quartile	Median	Upper quartile	90th percentile
Call_SVR (%)	-1.54	2.34	-4.47	-2.63	-1.27	-0.35	0.82
Put_SVR (%)	-1.56	1.55	-3.59	-2.33	-1.33	-0.67	0.02
AUTO	-0.01	0.10	-0.14	-0.08	-0.01	0.05	0.11
CASH	0.17	0.18	0.02	0.04	0.10	0.24	0.44
CFV	-0.05	0.36	-0.02	0.00	0.00	0.00	0.00
DISP (%)	13.82	44.30	0.80	1.47	3.22	8.25	24.13
HV-IV	0.02	0.09	-0.07	-0.03	0.01	0.05	0.11
ISSUE_1Y	-0.41	1.20	-1.14	-0.32	-0.01	0.02	0.05
IVOL	0.02	0.01	0.01	0.01	0.02	0.02	0.03
LNPRICE	3.71	0.71	2.78	3.28	3.76	4.17	4.54
LOG_AMIHUD	-8.36	1.53	-10.20	-9.39	-8.52	-7.44	-6.27
PROFIT	0.28	0.42	0.02	0.14	0.24	0.36	0.55
TEF	0.00	0.13	-0.14	-0.03	0.01	0.06	0.11

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

	Call_SVR	Put_SVR	IVOL	HV-IV	LOG_AMIHUD	OSPREAD	AUTO	DISP
Call_SVR	1.00	0.57	-0.09	0.00	-0.12	-0.04	-0.01	-0.04
Put_SVR		1.00	-0.12	-0.03	-0.17	-0.06	-0.01	-0.06
IVOL			1.00	0.12	0.41	0.08	0.11	0.24
HV-IV				1.00	0.00	-0.03	-0.03	0.03
LOG_AMIHUD					1.00	0.50	0.06	0.16
OSPREAD						1.00	0.00	0.06
AUTO							1.00	0.05
DISP								1.00

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

Panel A reports the average monthly returns to the delta-hedged option portfolios sorted by Call_SVR (Put_SVR). At each end of 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 2.1. 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 high-low return spread (i.e., the difference between the returns of the top and bottom quintile portfolios). At the end of each month, we use three weighting schemes when computing the average return of a portfolio: equal weight (EW), weight by the market capitalization of the underlying stock (Stock-VW), and weight by the market value of option open interest (Option-VW). We also adjust the average returns using a seven-factor model and report the corresponding alphas. Panel B examines the influence of transaction costs (bid-ask spread) on the profitability of our option portfolio strategy based on ML textual predictors. We form portfolio sorts just as in Panel A but take into consideration transaction costs when computing the realized returns. Each row corresponds to a given level of effective spread ESPR (e.g., equals to 50% of the quoted bid-ask spread QSPR). We report the (5-1) portfolio return spreads after accounting for the assumed option transaction costs. The weighting scheme in Panel B is equal-weighted. All returns are expressed in percentage. The sample period is from February 1996 to December 2018. 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: Average returns and Alphas for Portfolios Sorted by SVR

		Low	**2**	**3**	**4**	**High**	**H-L**	**Alpha**
Call Options	EW	-3.22 (-20.44)	-3.27 (-23.06)	-2.85 (-21.32)	-2.01 (-15.09)	-1.75 (-12.04)	1.47*** (12.47)	1.48*** (11.17)
	Stock-VW	-3.04 (-20.34)	-3.13 (-22.58)	-2.77 (-21.31)	-1.92 (-14.56)	-1.62 (-11.36)	1.42*** (12.55)	1.43*** (11.07)
	Option-VW	-3.00 (-17.92)	-3.30 (-18.53)	-2.76 (-19.61)	-1.55 (-10.93)	-1.05 (-5.00)	1.95*** (8.74)	1.84*** (7.94)
		Low	**2**	**3**	**4**	**High**	**H-L**	**Alpha**
Put Options	EW	-2.40 (-19.81)	-2.20 (-20.66)	-2.05 (-18.67)	-1.78 (-14.53)	-1.53 (-11.66)	0.87*** (13.58)	0.86*** (10.16)
	Stock-VW	-2.26 (-18.94)	-2.09 (-19.82)	-1.95 (-18.49)	-1.68 (-13.91)	-1.43 (-10.84)	0.83*** (12.91)	0.83*** (9.68)
	Option-VW	-2.35 (-15.03)	-2.05 (-15.12)	-1.97 (-17.26)	-1.46 (-9.69)	-1.10 (-5.64)	1.25*** (7.79)	1.21*** (7.49)

Panel B: Portfolio Performances after Accounting for Transaction Costs

			ESPR/QSPR		
	No Cost	**25%**	**50%**	**75%**	**100%**
Call Options	1.47*** (12.47)	1.41*** (12.11)	1.34*** (11.70)	1.28*** (11.25)	1.21*** (10.77)
Put Options	0.87*** (13.58)	0.80*** (12.56)	0.73*** (11.42)	0.65*** (10.18)	0.58*** (8.87)

Table 3: Dependent Double Sorts and Fama-MacBeth Regressions

In Panel A of this table, we investigate whether several stock or option characteristics can explain the effect of ML textual predictors using dependent double sorts. We first sort all options into quintiles based on a given control variable such as idiosyncratic volatility (IVOL), volatility deviation (HV – IV), Amihud illiquidity measure (LOG_AMIHUD), options' bid-ask spread, gamma, and vega (OSPREAD, Gamma, Vega), autocorrelation (AUTO), analyst earnings forecast dispersion (DISP), cash flow variance (CFV), cash-to-assets ratio (CASH), 1-year new issues (ISSUE_1Y), log of the underlying stock price (LNPRICE), profitability (PROFIT), total external finance (TEF). Then, within each quintile we further sort the options into five portfolios based on ML-based textual predictors. Finally, we average returns for each textual predictor quintile across the five groups sorted by the control variable, yielding five control-variable adjusted quintile returns. We report the baseline results based on univariate sort (i.e., the average difference in the equal-weighted monthly returns of the top and bottom quintile portfolios sorted by SVR textual predictors) in the first row, followed by the corresponding results after controlling for the variable labeled in each subsequent row. Panel B reports the Fama-Macbeth cross-sectional regression results of delta-hedged equity option returns on SVR textual predictors, Call_SVR (Put_SVR). Detailed descriptions of Call_SVR (Put_SVR) and their constructions are provided in Section 2.2. The constructions of control variables are described in Table 1. The sample period is from February 1996 to December 2018. 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: Portfolio Return Spread based on SVR Predictors after Controlling Stock or Option Characteristics

	Call Options	Put Options
SVR	1.47*** (12.47)	0.87*** (13.58)
IVOL	1.13*** (11.37)	0.60*** (9.59)
HV-IV	1.30*** (11.91)	0.81*** (13.55)
LOG_AMIHUD	1.32*** (10.62)	0.58*** (7.55)
OSPREAD	1.45*** (12.32)	0.82*** (12.07)
AUTO	1.42*** (12.27)	0.84*** (13.54)
Gamma	1.40*** (12.14)	0.79*** (13.63)
Vega	1.48*** (12.79)	0.81*** (12.25)
DISP	1.27*** (11.38)	0.68*** (10.64)
CFV	1.28*** (11.03)	0.74*** (10.13)
CASH	1.29*** (11.24)	0.77*** (10.80)
ISSUE_1Y	1.32*** (12.45)	0.77*** (12.60)
LNPRICE	1.16*** (10.38)	0.59*** (10.69)
PROFIT	1.26*** (10.39)	0.73*** (10.88)
TEF	1.32*** (11.10)	0.78*** (11.96)

Panel B: Fama-Macbeth Regressions

	Call Options			Put Options		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
SVR	0.465*** (11.90)	0.307*** (8.69)	0.259*** (7.06)	0.290*** (13.92)	0.142*** (7.12)	0.118*** (4.32)
IVOL		-0.764*** (-13.27)	-0.445*** (-8.30)		-0.624*** (-15.78)	-0.301*** (-6.77)
HV-IV		0.484*** (8.12)	0.514*** (9.38)		0.434*** (9.76)	0.446*** (10.35)
LOG_AMIHUD		-0.357*** (-11.06)	0.012 (0.31)		-0.355*** (-13.08)	0.021 (0.51)
OSPREAD			-0.066* (-1.81)			-0.118*** (-2.90)
AUTO			0.026 (1.03)			0.031 (1.55)
Gamma			0.471*** (5.69)			0.598*** (8.32)
Vega			-0.308*** (-14.22)			0.042 (1.58)
Stock Characteristics	No	No	Yes	No	No	Yes
Adj. R^2 (%)	1.615	14.143	22.536	1.177	13.687	21.951

Table 4: Option Portfolios Sorted by Alternative Machine Learning Textual Predictors

Panel A reports average monthly returns of the delta-hedged option portfolios sorted by machine learning (ML) textual predictors trained by alternative machine learning algorithms. The row "ENET"/"RF"/"MLP" reports portfolio sorting results by textual predictors extracted based on elastic net, random forest, and neural networks. Detailed descriptions of these predictors are provided in Section 3.2.1. Panel B reports the average monthly returns of the delta-hedged option portfolios sorted by ML textual predictors trained by using alternative word constructions. The row "Bigram"/"Trigram"/"Fivegram" reports portfolio sorting results based on ML textual predictors extracted based on different word features to train the model, including bigram, trigram, and fivegram. Detailed descriptions of these predictors are provided in Section II.B.2. All returns are expressed in percentage. The sample period is from February 1996 to December 2018. 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: Portfolios Sorted by ML Textual Predictors using Alternative ML Algorithms

		Low	**2**	**3**	**4**	**High**	**H-L**	**Alphas**
Call Options	SVR	-3.22 (-20.44)	-3.27 (-23.06)	-2.85 (-21.32)	-2.01 (-15.09)	-1.75 (-12.04)	1.47*** (12.47)	1.48*** (11.17)
	ENET	-3.21 (-21.62)	-2.82 (-19.52)	-2.49 (-20.86)	-2.31 (-16.06)	-2.27 (-15.87)	0.94*** (9.39)	0.93*** (7.86)
	RF	-3.15 (-21.90)	-2.96 (-20.61)	-2.70 (-20.12)	-2.23 (-16.54)	-2.06 (-14.52)	1.09*** (11.41)	1.02*** (9.00)
	MLP	-3.17 (-21.83)	-3.23 (-20.27)	-2.79 (-20.78)	-2.12 (-15.61)	-1.78 (-12.37)	1.38*** (12.17)	1.35*** (10.62)
		Low	**2**	**3**	**4**	**High**	**H-L**	**Alphas**
Put Options	SVR	-2.40 (-19.81)	-2.20 (-20.66)	-2.05 (-18.67)	-1.78 (-14.53)	-1.53 (-11.66)	0.87*** (13.58)	0.86*** (10.16)
	ENET	-2.31 (-18.47)	-2.06 (-18.57)	-1.98 (-16.92)	-1.81 (-15.19)	-1.80 (-15.23)	0.51*** (8.01)	0.47*** (5.86)
	RF	-2.34 (-17.96)	-2.12 (-18.92)	-1.99 (-16.94)	-1.79 (-15.80)	-1.71 (-14.53)	0.63*** (10.02)	0.55*** (6.84)
	MLP	-2.33 (-17.54)	-2.21 (-20.76)	-2.07 (-19.39)	-1.80 (-14.48)	-1.54 (-11.55)	0.78*** (9.48)	0.77*** (8.55)

Panel B: Portfolios Sorted by ML Textual Predictors under Alternative Word Constructions

		Low	**2**	**3**	**4**	**High**	**H-L**	**Alphas**
Call Options	Unigram	-3.22	-3.27	-2.85	-2.01	-1.75	1.47^{***}	1.48^{***}
		(-20.44)	(-23.06)	(-21.32)	(-15.09)	(-12.04)	(12.47)	(11.17)
	Bigram	-3.13	-3.33	-2.88	-2.12	-1.63	1.50^{***}	1.47^{***}
		(-21.10)	(-20.13)	(-20.78)	(-17.04)	(-11.45)	(11.77)	(13.79)
	Trigram	-3.22	-3.14	-2.97	-2.15	-1.62	1.60^{***}	1.64^{***}
		(-21.28)	(-20.66)	(-21.16)	(-15.70)	(-11.72)	(11.54)	(12.04)
	Fivegram	-3.58	-2.90	-2.57	-2.30	-1.88	1.69^{***}	1.69^{***}
		(-21.45)	(-20.12)	(-17.31)	(-18.58)	(-15.05)	(14.02)	(14.26)
		Low	**2**	**3**	**4**	**High**	**H-L**	**Alphas**
Put Options	Unigram	-2.40	-2.20	-2.05	-1.78	-1.53	0.87^{***}	0.86^{***}
		(-19.81)	(-20.66)	(-18.67)	(-14.53)	(-11.66)	(13.58)	(10.16)
	Bigram	-2.28	-2.24	-2.06	-1.84	-1.54	0.74^{***}	0.79^{***}
		(-18.65)	(-19.28)	(-20.24)	(-15.72)	(-9.84)	(7.66)	(7.98)
	Trigram	-2.29	-2.16	-2.08	-1.81	-1.61	0.68^{***}	0.74^{***}
		(-20.50)	(-18.43)	(-19.04)	(-14.89)	(-11.35)	(7.05)	(6.37)
	Fivegram	-2.40	-2.16	-2.01	-1.81	-1.58	0.82^{***}	0.80^{***}
		(-19.11)	(-17.85)	(-18.50)	(-15.20)	(-12.69)	(9.06)	(8.56)

Table 5: Alternative Option Samples and Return Construction Method

This table examines the robustness of the predictability by the SVR predictors in different option samples and using an alternative return construction method. We replicate the portfolio sorts in Table 2 using options with different moneyness and maturities. ITM represents in-the-money option, and OTM represents out-of-the-money option. 2.5-month ATM (3.5-month ATM are at-the-money options with 2.5-month (3.5-month) maturity. Daily-rebalance represents results using daily-rebalanced delta-hedged option return. The weighting scheme is equal-weighted. The sample period is from February 1996 to December 2018. 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: Alternative Samples and Return Construction - Call Options

	Low	2	3	4	High	H-L
ITM	-1.70 (-15.99)	-1.77 (-15.59)	-1.42 (-15.69)	-1.00 (-11.34)	-0.97 (-9.83)	0.74*** (9.51)
OTM	-4.04 (-20.02)	-4.15 (-23.06)	-3.69 (-21.27)	-2.83 (-16.49)	-2.26 (-10.61)	1.78*** (10.46)
2.5-month ATM	-2.40 (-15.36)	-2.37 (-16.59)	-2.21 (-18.81)	-1.34 (-8.69)	-1.04 (-6.88)	1.35*** (8.95)
3.5-month ATM	-2.23 (-16.32)	-2.22 (-16.69)	-1.88 (-16.58)	-1.34 (-10.41)	-1.21 (-9.09)	1.01*** (10.03)
Daily rebalancing	-0.98 (-9.50)	-0.94 (-9.54)	-0.70 (-6.61)	-0.41 (-3.81)	-0.29 (-2.34)	0.69*** (9.12)

Panel B: Alternative Samples and Return Construction - Put Options

	Low	2	3	4	High	H-L
ITM	-1.02 (-14.30)	-0.91 (-15.15)	-0.79 (-14.00)	-0.67 (-10.96)	-0.59 (-8.16)	0.43*** (9.02)
OTM	-3.28 (-16.26)	-3.01 (-19.50)	-2.85 (-14.80)	-2.55 (-12.91)	-2.23 (-10.46)	1.05*** (12.46)
2.5-month ATM	-1.55 (-12.74)	-1.39 (-11.64)	-1.35 (-11.94)	-1.08 (-11.20)	-0.79 (-4.38)	0.76*** (5.79)
3.5-month ATM	-1.26 (-11.58)	-1.22 (-12.45)	-1.09 (-11.03)	-0.97 (-8.75)	-0.90 (-9.19)	0.35*** (5.63)
Daily rebalancing	-0.79 (-7.59)	-0.67 (-6.18)	-0.58 (-6.41)	-0.46 (-4.46)	-0.40 (-3.90)	0.39*** (10.31)

Table 6: Topic Contribution for Feature Importance

Panel A and Panel B of this table list the proportion of important words from each topic category (labeled by the rows). The General Inquirer (GI) is a dictionary that reflects Charles Osgood's semantic differential findings regarding basic language universals. The Loughran and McDonald (LM) dictionary is the sentiment dictionary developed in Loughran and McDonald (2011). The Political dictionary is from Hassan et al. (2019). The Macroeconomic dictionary is based on Bloom (2014) and Bybee et al. (2020). Idiosyncratic Words are the remaining words that do not fall into the four pre-defined dictionaries. The sample period is from February 1996 to December 2018.

Panel A: Delta-hedged Call Option Feature Importance in Different Topics

Topic	SVR	Elastic Net	Random Forest	Neural Networks
GI Dictionary	37.40%	40.20%	40.60%	32.20%
LM Dictionary	32.6%	30.40%	32.20%	22.80%
Politics	4.60%	4.00%	6.0%	6.60%
Macroeconomics	8.00%	8.40%	7.80%	6.60%
Idiosyncratic Words	17.40%	17.00%	13.40%	31.80%
Total	100.00%	100.00%	100.00%	100.00%

Panel B: Delta-hedged Put Option Feature Importance in Different Topics

Topic	SVR	Elastic Net	Random Forest	Neural Networks
GI Dictionary	38.2%	44.60%	38.20%	28.40%
LM Dictionary	32.8%	27.60%	32.60%	24.60%
Politics	3.80%	6.40%	6.20%	5.20%
Macroeconomics	7.80%	5.00%	5.40%	7.60%
Idiosyncratic Words	17.40%	16.40%	17.60%	34.20%
Total	100.00%	100.00%	100.00%	100.00%

Table 7: Relationships between SVR Predictor and Sentiment-related Measures

Panel A reports Fama-Macbeth regression results of sentiment-related measures on the SVR predictor. GILM^1 is a proxy for the general opinions about a specific company extracted from the news media. GILM^2 is a measure of sentiment disagreement for each firm to reflect possible different opinions among the articles about a particular firm. Raven^1 and Raven^2 are sentiment level and disagreement measures from RavenPack News Analytics, respectively. Panel B reports the (5-1) portfolio return spreads generated by SVR predictors, sentiment-related measures, and SVR_GILM. SVR_GILM is an alternative SVR predictor that is trained using only sentences containing sentiment-related words in each article. The sample period is from February 1996 to December 2018. 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: Regressing Sentiment-related Measures on SVR Predictors

	Call				Put			
	GILM^1	GILM^2	Raven^1	Raven^2	GILM^1	GILM^2	Raven^1	Raven^2
SVR	0.010*** (5.96)	-0.024*** (-12.32)	0.052*** (4.82)	-0.040*** (-3.01)	0.012*** (5.81)	-0.034*** (-17.60)	0.060*** (5.19)	-0.052*** (-3.75)
Adj. R^2 (%)	0.596	1.911	0.244	0.243	0.712	2.985	0.140	0.283

Panel B: Portfolio Return Spreads Generated by SVR predictors, Sentiment-related Measures, and SVR_GILM predictors

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	SVR	GILM^1	GILM^2	Raven^1	Raven^2	SVR_GILM
Call	1.47*** (12.47)	0.51*** (5.54)	-0.62*** (-7.19)	0.68*** (7.54)	-0.60*** (-6.64)	0.12* (1.89)
Put	0.87*** (13.58)	0.41*** (5.48)	-0.50*** (-7.67)	0.59*** (8.92)	-0.51*** (-7.58)	0.18*** (3.38)

Panel C: Regressing Delta-hedged Option Returns on SVR Predictors with Sentiment-related Measures as Control Variables

	Call				Put			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SVR	0.460*** (12.00)	0.447*** (11.38)	0.454*** (8.68)	0.460*** (8.86)	0.289*** (13.91)	0.273*** (13.20)	0.283*** (9.80)	0.292*** (9.94)
GILM^1	0.163*** (6.05)				0.124*** (5.75)			
GILM^2		-0.188*** (-7.46)				-0.142*** (-7.16)		
Raven^1			0.248*** (7.89)				0.213*** (10.06)	
Raven^2				-0.232*** (-6.96)				-0.203*** (-8.48)
Adj.R^2 (%)	2.000		3.021	2.624	1.531	1.596	2.469	2.235

Table 8: Potential Economic Channels

Panel A examines potential economic channels underlying the predictive power of SVR predictors for delta-hedged option returns. The dependent variable is the delta-hedged option returns, and the independent variables are SVR predictors and proxies for various potential economic channels. VOIV is the standard deviation of option implied volatility over the trading days within a given month, and the implied volatility is the at-the-money implied volatility obtained from Volatility Surface provided by OptionMetrics IvyDB database. VORV is the standard deviation of the daily realized volatility of a given month, and the realized volatility is calculated with five-minute log returns provided by TAQ. MFIS (resp. MFIK) is the model-free option-implied skewness (resp. kurtosis). OM is the total market value of options on a stock scaled by the stock market value. OIB is the option demand pressure, calculated as the difference between buyer and seller-initiated option trading volume scaled by the total option trading volume. Option trading volume for firm *i* in month *t* is the sum of trading volumes of all tradable options written on firm i in month t. The sample period is from February 1996 to December 2018. To adjust for serial correlations, robust Newey and West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

			Panel A: Call Options			
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
SVR	0.449*** (11.73)	0.445*** (11.41)	0.433*** (11.61)	0.436*** (11.73)	0.448*** (11.87)	0.343*** (12.09)
VOIV	-0.252*** (-6.49)					
VORV		-0.242*** (-6.51)				
MFIS			-0.542*** (-15.31)			
MFIK				0.510*** (11.05)		
OM					-0.358*** (-8.20)	
OIB						-0.156*** (-7.56)
Adj.R2 (%)	8.129	5.443	4.049	4.379	4.344	1.974

			Panel B: Put Options			
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
SVR	0.285*** (13.87)	0.281*** (12.95)	0.272*** (13.41)	0.270*** (13.42)	0.255*** (12.14)	0.241*** (11.50)
VOIV	-0.218*** (-7.20)					
VORV		-0.213*** (-6.62)				
MFIS			-0.260*** (-11.01)			
MFIK				0.389*** (12.48)		
OM					-0.538*** (-16.12)	
OIB						0.065*** (3.49)
Adj.R^2 (%)	8.725	5.930	3.459	4.542	6.991	2.417

Table 9: Information Content of SVR Predictor

In this table, we investigate the information content of the SVR predictor. We run predictive regression with future implied volatility changes over the next month, earnings surprises in the next month, or stock excess over the next month returns as the dependent variable The key independent variable here is the SVR predictor. When running regressions for future implied volatility changes, we add their lag values ($Lag\_\Delta CVOL$ or $Lag\_\Delta PVOL$) as control variables. SUE is the standardized unexpected earnings excluding "special items" as in Livnat and Mendenhall (2006). Stock return is the excess stock return over the next month. The sample period is from February 1996 to December 2018. To adjust for serial correlations, robust Newey and West (1987) t-statistics are reported in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Call			Put		
	$\Delta CVOL$	SUE	Stock Return	$\Delta PVOL$	SUE	Stock Return
SVR	0.559*** (7.13)	0.004 (1.04)	0.075 (1.44)	0.194*** (2.87)	0.004 (1.53)	0.073 (1.38)
$Lag\_\Delta CVOL$ ($\Delta PVOL$)	-5.008*** (-27.54)			-4.894*** (-24.50)		
Adj.R^2 (%)	11.111	0.089	0.494	11.064	0.039	0.433

Table A1: Feature Importance with Different Topics and Dictionaries

This table lists the word feature importance from the support vector regression (SVR) model on average over time. 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 twenty words which appear most often over time for both call and put options. The sample period is from February 1996 to December 2018.

Topic	Delta-hedged Call Option Return	Delta-hedged Put Option Return
GI Dictionary	excessive, aggregate, parasite, endear, tolerable, unpopular, discretion, grumble, purity, heartily, foolish, simplistic, stark, vitality, conscience, reactive, intelligent, flaw, godliness, capability	fiery, animosity, learner, upright, glorify, massacre, obstinate, capitalize, stability, transgress, enthusiasm, lowly, evict, guilt, discretion, fashionable, fullness, veto, devious, peril
LM Dictionary	exposing, overstate, miscalculating, acquitting, underestimate, forgers, perfects, defamatory, overshadowed, purported, cancelled, exacerbated, annulments, revolutionizing, ridiculed, exculpatory, inaction, deceived, demotion, dysfunction	irreparable, mishandling, forecloses, overshadowed, unremedied, exclusive, abolishes, inadequacies, provoked, revocations, summons, breakthrough, confesses, noncompliances, overlooked, distressed, insufficiently, stressed, expropriates, inappropriately
Politics	caviar, apparatchik, white house, command, floor, autocracy, trojan, wets, taliban, dynasty, political, roots, libertarianism, rotherham, bureaucracy, republican, legislature, humanism, religion, congress	polling, pravda, election, bien, senate, levellers, proportional, governor, president, ombudsman, belli, white house, mission, patrician, political, government, reform, obama, court, religion, republican
Macroeconomics	group president, deadlock, price earnings, aeronautic defence, definitive agreement, secondary offer, aeronautic, nonprofit group, noncompetly, operate loss, commodity market, stock market, nasdaq, london interbank, bilateral trade, charity, export country, omnicom group, report conclude, ambitious plan	rose slight, creditcard issuer, credit agreement, transportation safety, complexity, issue bond, kim jong, grid, warn sign, traffic safety, york cotton, misappropriated, aid virus, fargo, redeemable, commodity oil, retiree, stanley capital, criminal trial
Idiosyncratic Words	exception, interpleader, contract, despair, lawyer, internet, electronic, fighting, bioteh, indemnitee, shot, properly, contravention, statutorily, tech, admissions, hospital, aerospace, entailing, pledgors	crowd, willfulness, blood, passenger, nurse, increase, obligee, offered, subrogated, iphone, counsel, indemnifications, necessitated, shopper, mandamus, aforedescribed, rebate, mediation, supersedeas, advertising, truck

Table A2: Important Words to Understand the SVR Textual Predictors

In Panel A, we construct synthetic bigrams using both important firm-idiosyncratic words and sentiment words selected by SVR. To make the bigram more interpretable, we first list the firm-idiosyncratic word followed by the sentiment word. In Panel B, we list some fivegram words that have the highest average feature importance (i.e., those with largest average absolute value of coefficients under support vector regression model). For abbreviation, we list the first eighty words which appear most often over time for both call and put options. The sample period is from February 1996 to December 2018.

Panel A: Important Features based on Bigrams using Sentiment and Firm-idiosyncratic Words
said-good, fell-lost, department-justice, comment-declined, said-never, financial-crisis, electronics-success, added-gained, said-able, shares-lost, said-justice, general-attorney, said-might, said-late, shares-closed, said-contract, still-good, news-good, would-risk, corporate-lawyers, said-risk, said-could, advice-lawyers, said-best, last-strong, said-great, said-investigation, shares-gained, would-warned, cents-lost, said-probably, said-confident, would-able, years-could, first-lost, another-could, many-could, many-never, advice-defendant, billion-could, results-strong, companies-could, market-could, says-good, earnings-strong, million-loss, price-could, think-good, chance-good, companies-good, case-could, said-better, would-probably, said-arrest, nasdaq-gained, fell-declined, ground-lost, advice-counsel, would-concerns, federal-court, team-success, much-could, executive-fraud, means-lawyers, basis-lawyers, earnings-gains, billion-nearly, advice-fraud, conduct-argue, hosts-suspects, business-could, said-abolished, said-stronger, annually-illegal, actually-negative, started-assault, would-restructuring, another-weaker, amazon-could, largest-assault, fallen-doubt, statement-confident, consumer-success
Panel B: Important Features based on Fivegrams
be-determine-difference-price-face, chief-executive-say-conference-call, figure-sale-million-firm-distributor, cent-share-compare-net-loss, chairman-chief-executive-officer-say, boxoffice-figure-sale-million-firm, continue-follow-fashion-lifestyle-coverage, chief-executive-say-company-be, executive-vice-president-chief-financial, be-issue-advance-decline-unchanged, call-gain-volume-contract-compare, chief-executive-officer-say-company, excellent-very-good-good-satisfactory, change-discounter-million-year-early, compare-net-loss-cent-share, foreign-noncompetitively-auction-price-rate, advance-issue-top-decliner-board, apple-chief-executive-steve-job, executive-say-conference-call-analyst, comparable-sale-store-change-change, possible-be-payable-earning-share, be-name-executive-vice-president, chief-executive-chief-financial-officer, decline-issue-finish-well-ahead, have-be-appoint-nonexecutive-director, agree-acquire-banking-concern-stock, be-name-president-chief-executive, compare-outstanding-contract-gain-cent, earning-late-beat-wall-street, easily-top-wall-street-forecast, executive-tell-analyst-conference-call, invite-question-computerbase-technology-tech, be-close-replace-samestore-sale, be-develop-show-positive-result, be-measure-bad-performer-drop, be-much-strong-analyst-be, capital-gain-loss-dividend-theoretical, change-million-year-early-year, computer-lose-international-business-machine, concern-report-fiscal-firstquarter-result, concern-report-fiscal-thirdquarter-result, executive-vice-president-sale-marketing, fund-troubled-asset-relief-program, government-trouble-asset-relief-program, loss-cent-share-compare-profit