WORKING PAPER · NO. 2019-69

Predicting Returns with Text Data

Zheng Tracy Ke, Brian Kelly, Dacheng Xiu

MAY 2019

Becker
Friedman
Institute
FOR ECONOMICS AT UCHICAGO

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

Predicting Returns with Text Data*

Zheng Tracy Ke
Department of Statistics
Harvard University

Bryan Kelly
Yale University, AQR Capital
Management, and NBER

Dacheng Xiu
Booth School of Business
University of Chicago

May 14, 2019

Abstract

We introduce a new text-mining methodology that extracts sentiment information from news articles to predict asset returns. Unlike more common sentiment scores used for stock return prediction (e.g., those sold by commercial vendors or built with dictionary-based methods), our supervised learning framework constructs a sentiment score that is specifically adapted to the problem of return prediction. Our method proceeds in three steps: 1) isolating a list of sentiment terms via predictive screening, 2) assigning sentiment weights to these words via topic modeling, and 3) aggregating terms into an article-level sentiment score via penalized likelihood. We derive theoretical guarantees on the accuracy of estimates from our model with minimal assumptions. In our empirical analysis, we text-mine one of the most actively monitored streams of news articles in the financial system—the *Dow Jones Newswires*—and show that our supervised sentiment model excels at extracting return-predictive signals in this context.

Key words: Text Mining, Machine Learning, Return Predictability, Sentiment Analysis, Screening, Topic Modeling, Penalized Likelihood

*We thank Kangying Zhou and Mingye Yin for excellent research assistance. We benefited from discussions with Rob Engle, Timothy Loughran, as well as seminar and conference participants at the New York University, Yale University, Ohio State University, Hong Kong University of Science and Technology, University of Zurich, AQR, NBER Conference on Big Data: Long-Term Implications for Financial Markets and Firms, Market Microstructure and High Frequency Data Conference at the University of Chicago, the Panel Data Forecasting Conference at USC Dornsife Institute, and the 8th Annual Workshop on Applied Econometrics and Data Science at the University of Liverpool. We gratefully acknowledge the computing support from the Research Computing Center at the University of Chicago. AQR Capital Management is a global investment management firm, which may or may not apply similar investment techniques or methods of analysis as described herein.

1 Introduction

Advances in computing power have made it increasingly practicable to exploit large and often unstructured data sources such as text, audio, and video for scientific analysis. In the social sciences, textual data is the fastest growing data form in academic research. The numerical representation of text as data for statistical analysis is, in principle, ultra-high dimensional. Empirical research seeking to exploit its potential richness must also confront its dimensionality challenge. Machine learning offers a toolkit for tackling the high-dimensional statistical problem of extracting meaning from text for explanatory and predictive analysis.

While the natural language processing and machine learning literature is growing increasingly sophisticated in its ability to model the subtle and complex nature of verbal communication, usage of textual analysis in empirical finance is in its infancy. Text has most commonly been used in finance to study the "sentiment" of a given document, and this sentiment has been most frequently measured by weighting terms based on a pre-specified sentiment dictionary (e.g., the Harvard-IV psychosocial dictionary) and summing these weights into document-level sentiment scores. Document sentiment scores are then used in a secondary statistical model for investigating a financial research question such as "how do asset returns associate with media sentiment?"

Highly influential studies in this area include Tetlock (2007) and Loughran and McDonald (2011). These papers manage the dimensionality challenge by restricting their analysis to words in pre-existing sentiment dictionaries and using ad hoc word-weighting schemes. This approach has the great advantage that it allows researchers to make progress on understanding certain aspects of the data without taking on the (often onerous) task of estimating a model for a new text corpus from scratch. But it is akin to using model estimates from a past study to construct fitted values in a new collection of documents being analyzed.

The goal of this paper is to demonstrate how basic machine learning techniques can be used to understand the sentimental structure of a text corpus without relying on pre-existing dictionaries. The method we suggest has three main virtues. The first is simplicity—it requires only standard econometric techniques like correlation analysis and maximum likelihood estimation. Second, our method requires minimal computing power—it can be run with a laptop computer in a matter of minutes for text corpora with millions of documents. Third, and most importantly, it allows the researcher to construct a sentiment scoring model that is specifically adapted to the context of the dataset at hand. This frees the researcher from relying on a pre-existing sentiment dictionary that was originally designed for different purposes. A central hurdle to testing theories of information economics is the difficulty of quantifying information. Our estimator is a sophisticated (yet easy to use) tool for measuring the information content of text documents that opens new lines of research into empirical information economics.

Our empirical analysis revisits perhaps the most commonly studied text-based research question in finance, the extent to which business news explains and predicts observed asset price variation. We analyze the machine text feed and archive database of the *Dow Jones Newswires*, which is widely subscribed and closely monitored by market participants, and available over a 38-year time span.

Its articles are time-stamped and tagged with identifiers of firms to which an article pertains. Using these identifiers, we match articles with stock return data from CRSP in order to model stock-level return behavior as a function of a Newswire content. We then learn the sentiment scoring model from the joint behavior of article text and stock returns. This is a key aspect of our approach—learning the sentiment scoring model from the joint behavior of article text and stock returns—rather than taking sentiment scores off the shelf.

Our estimated model isolates an interpretable and intuitive list of positive and negative sentiment words and their associated sentiment scores. We demonstrate the predictive capacity of our model through a simple trading strategy that buys assets with positive recent news sentiment and sells assets with negative sentiment. The portfolio based on our model delivers excellent risk-adjusted out-of-sample returns, and outperforms a similar strategy based on scores from RavenPack (the industry-leading commercial vendor of financial news sentiment scores).

We compare the price impact of “fresh” versus “stale” news by devising a measure of article novelty (its cosine distance from the language in articles about the same stock over the preceding week). While even stale news significantly correlates with future price changes, the price effect in the first day after news arrival is 70% larger for fresh news. And while the effects of stale news are fully reflected in prices within two days of arrival, it takes four days for fresh news to be completely assimilated. Likewise, we study the differences in price assimilation of news based on stock attributes. We find that price responses to news are roughly four times as large for smaller (below NYSE median) and more volatile (above median) stocks, and that it takes roughly twice as long for their news to be fully reflected in prices.

Unlike many commercial platforms or most deep learning approaches which amount to black boxes for their users, the supervised learning approach we propose is entirely “white box.” We call our procedure SSESTM (pronounced “system,” for Supervised Sentiment Extraction via Screening and Topic Modeling). The model consists of three parts, and basic machine learning strategies play a central role in each.

The first step isolates the most relevant features from a very large vocabulary of terms. The vocabulary is derived from the bag-of-words representation of each document as a vector of term counts. We take a variable selection approach to extracting a comparatively small number of terms that are likely to be informative for asset returns. In this estimation step, variable selection via correlation *screening* is the necessary machine learning ingredient for fast and simple estimation of our reduced-dimension sentiment term list. The idea behind screening is to find individual terms—positive or negative—that most frequently coincide with returns of the same sign. It is a natural alternative to regression and other common dimension reduction techniques (such as principal components analysis) which behave poorly when confronted with the high dimensionality and sparsity challenges of text data.

The second step is to assign term-specific sentiment weights based on their individual relevance for the prediction task. Text data is typically well approximated by Zipf’s law, which predicts a small number of very high-frequency terms and a very large number of low-frequency terms. While existing finance literature recognizes the importance of accounting for vast differences in term frequencies when assigning sentiment weights, the ultimate choice of weights has typically been ad hoc (e.g., via term frequency-inverse document frequency weighting, or "tf-idf"). We instead use a likelihood-based, or "generative," model to account for the extreme skewness in term frequencies. The specific machine learning tool we apply in this component is a supervised topic model. For the sake of simplicity and computational ease, and because it is well adapted to our purposes, we opt for a model with only two topics—one that describes the frequency distribution of positive sentiment terms, and one for negative sentiment words.

Finally, we use the estimated topic model to assign an article-level sentiment score. When aggregating to an article score, we use the internally consistent likelihood structure of the model to account for the severe heterogeneity in both the frequency of words as well as their sentiment weights. To robustify the model, we design a penalized maximum likelihood estimator with a single unknown parameter to estimate for each article. A Bayesian interpretation of the penalization is to impose a Beta-distributed prior on the sentiment score that centers at 1/2. That is, our estimation begins from the prior that an article is sentiment neutral.

We establish the theoretical underpinnings of the SSESTM algorithm. In particular, we shed light on its biases (if any) and statistical efficiency, and characterize how these properties depend on the length of the dictionary, the number of news articles, and the average number of words per article.

This paper contributes to a nascent literature using textual analysis via machine learning for financial research. Most prior work using text as data for finance and accounting research does little, if any, direct statistical analysis of text. In perhaps the earliest work on text mining for return prediction, Cowles (1933) manually reads and classifies editorials of *The Wall Street Journal* as bullish, bearish, or neutral. He finds that a trading strategy that follows editor recommendations destroys value in the 1902-1929 sample. More recent research relies largely on sentiment dictionaries (see Loughran and Mcdonald, 2016, for a review). These studies generally find that dictionary-based news sentiment scores are statistically significant predictors for future returns, though the economic magnitudes tend to be small. For example, applying the Harvard-IV psychosocial dictionary to a subset of articles from *The Wall Street Journal*, Tetlock (2007) finds that a one standard deviation increase in pessimism predicts an 8.1 basis point decline in the Dow Jones Industrial Average the following day (this is in-sample). Loughran and McDonald (2011) create a new sentiment dictionary specifically designed for the context of finance. They analyze 10-K filings and find that sentiment scores from their dictionary have a higher correlation with filing returns than scores based on Harvard-IV. They do not, however, explore predictive performance or portfolio choice. In contrast with this literature, we develop a machine learning method to build context-specific sentiment scores. We construct and evaluate the performance of trading strategies that exploit our sentiment estimates, and find large economic gains, particularly out-of-sample.

A few exceptions in the finance literature use machine learning to analyze text, and are surveyed in Gentzkow et al. (forthcoming). Using a Naïve Bayes approach, Antweiler and Frank (2005) find that internet stock messages posted on Yahoo Finance and Raging Bull for about 45 companies help predict market volatility, and the effect on stock returns is statistically significant but economically small. Other related work includes Li (2010), Jegadeesh and Wu (2013), and Huang et al. (2014). Manela and Moreira (2017) use support vector regression to relate frontpage text of *The Wall Street Journal* to the VIX volatility index. As Loughran and Mcdonald (2016) note, Naïve Bayes involves thousands of unpublished rules and filters to measure the context of documents, and hence is opaque and difficult to replicate. Lack of transparency is a research limitation of machine learning methods more generally. In contrast, our model is generative, transparent, tractable, and accompanied by theoretical guarantees. Our method is closer to modern text mining algorithms in computer science and machine learning, such as latent Dirichlet allocation (LDA, Blei et al., 2003) and its descendants, and vector representations of text such as word2vec (Mikolov et al., 2013). The key distinction between our model and many such machine learning approaches is that our method is supervised and thus customizable to specific prediction tasks. In this vein, our model is most similar to Gentzkow et al. (2019), who develop a supervised machine learning approach to study partisanship in congressional speech.

Finally, our research relates more broadly to a burgeoning strand of literature that applies machine learning techniques to asset pricing problems. In particular, Gu et al. (2018) review a suite of machine learning tools for return prediction using well established numerical features from the finance literature.[1] They find that some of the best performing numerical predictors are technical indicators, such as momentum and reversal patterns in stock prices. Our paper uses alternative data—news text—whose dimensionality vastly exceeds that used for return prediction in past work. And, unlike technical indicators that are difficult to interpret, the features in our analysis are counts of words, and are thus interpretable.

The rest of the paper is organized as follows. In Section 2, we set up the model and present our methodology. Section 3 conducts the empirical analysis. Section 4 concludes. The appendix contains the statistical theory, mathematical proofs, and Monte Carlo simulations.

2 Methodology

To establish notation, consider a collection of n news articles and a dictionary of m words. We record the word (or phrase) counts of the i^{th} article in a vector $d_i \in \mathbb{R}_+^m$, so that $d_{i,j}$ is the number of time word j occurs in article i. In matrix form, this is an $n \times m$ document-term matrix, $D = [d_1, ..., d_n]'$. We occasionally work with a subset of columns from D, where the indices of columns included in the subset are listed in the set S. We denote the corresponding submatrix as $D_{\cdot,[S]}$. We then use $d_{i,[S]}$ to denote the row vector corresponding to the i^{th} row of $D_{\cdot,[S]}$.

Articles are tagged with the identifiers of stocks mentioned in the articles. For simplicity, we study articles that correspond to a single stock,[2] and we label article i with the associated stock return (or its idiosyncratic component), y_i, on the publication date of the article.

[1] Other examples include Freyberger et al. (2017), Kozak et al. (2017), Kelly et al. (2017), and Feng et al. (2017).

[2] While this assumption is a limitation of our approach, the large majority of articles in our sample are tagged to a single firm. In general, however, it would be an advantage to handle articles about multiple firms. For instance, Apple and Samsung are competitors in the cellphone market, and there are news articles that draw a comparison between them. In this case, the sentiment model requires more complexity, and we leave such extensions for future work.

2.1 Model Setup

We assume each article possesses a sentiment score $p_i \in [0,1]$; when $p_i = 1$, the article sentiment is maximally positive, and when $p_i = 0$, it is maximally negative. Furthermore, we assume that p_i serves as a sufficient statistic for the influence of the article on the stock return. That is,

$$d_i \text{ and } y_i \text{ are independent given } p_i. \tag{1}$$

Along with the conditional independence assumption, we need two additional components to fully specify the data generating process. One governs the distribution of the stock return y_i given p_i, and the other governs the article word count vector d_i given p_i.

For the conditional return distribution, we assume

$$\mathbb{P}\big(\text{sgn}(y_i) = 1\big) = g(p_i), \text{ for a monotone increasing function } g(\cdot), \tag{2}$$

where sgn(x) is the sign function that returns 1 if $x > 0$ and 0 otherwise. Intuitively, this assumption states that the higher the sentiment score, the higher the probability of realizing a positive return. Note that this modeling assumption is rather weak—we do not need to specify the full distribution of y_i or the particular form of $g(\cdot)$ to establish our theoretical guarantees below.

We now turn to the conditional distribution of word counts in an article. We assume the dictionary has a partition:

$$\{1, 2, \ldots, m\} = S \cup N, \tag{3}$$

where S is the index set of sentiment-charged words, N is the index set of sentiment-neutral words, and $\{1, \ldots, m\}$ is the set of indices for all words in the dictionary (S and N have dimensions $|S|$ and $m - |S|$, respectively). Likewise, $d_{i,[S]}$ and $d_{i,[N]}$ are the corresponding subvectors of d_i and contain counts of sentiment-charged and sentiment-neutral words, respectively.

We assume that $d_{i,[S]}$ and $d_{i,[N]}$ are independent of each other. The distribution of sentiment-neutral counts, $d_{i,[N]}$, is essentially a nuisance, and due to its independence from the vector of interest, $d_{i,[S]}$, it suffices for our purposes to leave $d_{i,[N]}$ unmodeled.[3]

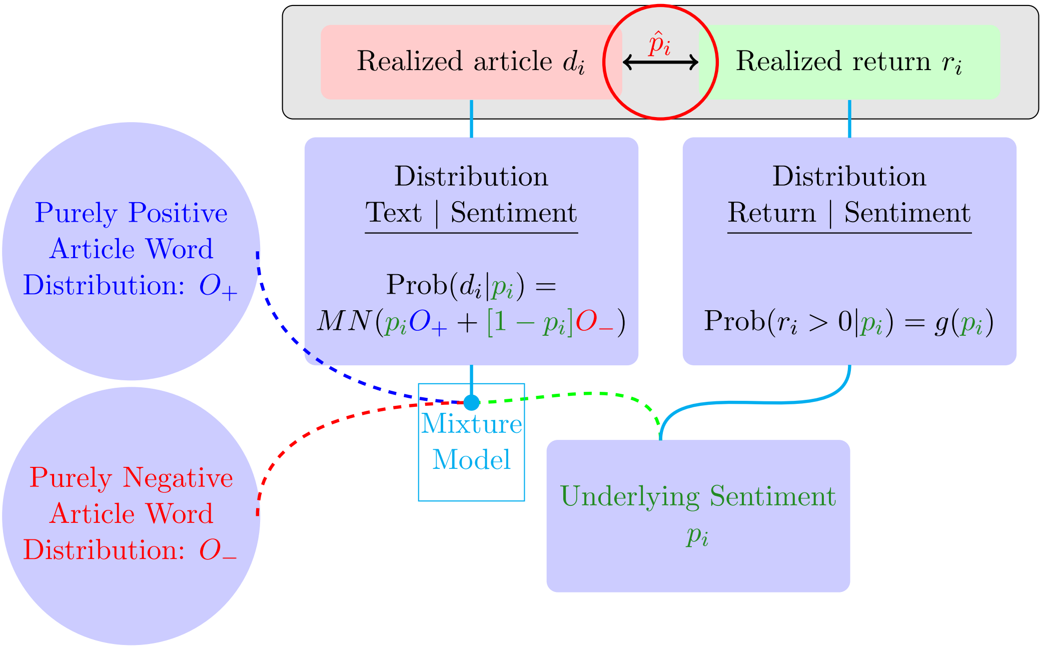
We assume that sentiment-charged word counts, $d_{i,[S]}$, are generated by a mixture multinomial distribution of the form

$$d_{i,[S]} \sim \text{Multinomial}\Big(s_i,\ p_i O_+ + (1 - p_i) O_-\Big), \tag{4}$$

where s_i is the total count of sentiment-charged words in article i, which determines the scale of the multinomial. Next, we model the probabilities of individual word counts with a two-topic mixture model. O_+ is a probability distribution over words—it is an $|S|$-vector of non-negative entries with unit ℓ^1-norm. O_+ is a "positive sentiment topic," and describes expected counts for words in a maximally positive sentiment article (one for which $p_i = 1$). Likewise, O_- is a "negative sentiment

[3] We may further model sentiment-neutral counts, $d_{i,[N]}$, using a standard K-topic model (Hofmann, 1999; Blei et al., 2003). This is, however, unnecessary in our setting due to our focus on sentiment extraction.

Figure 1: Model Diagram



Note: Illustration of model structure.

topic" that describes the distribution of word probabilities in maximally negative articles (those for which $p_i = 0$). At intermediate values of sentiment $0 < p_i < 1$, word probabilities are a convex combination of those from the positive and negative sentiment topics. A word j is a "positive word" if the j^{th} entry of $(O_+ - O_-)$ is positive, i.e., this word has a larger weight in the positive sentiment topic than in the negative sentiment topic. Similarly, a word j is a "negative word" if the j^{th} entry of $(O_+ - O_-)$ is negative.

Figure 1 provides a visualization of the model's structure. The data available to infer sentiment are in the box at the top of the diagram, and include not only the realized document text, but also the realized event return. The important feature of this model is that, for a given event i, the distribution of sentiment word counts and the distribution of returns are linked through the common parameter, p_i. Returns supervise the estimation and help identify which words are assigned to the positive versus negative topic. A higher p_i maps monotonically into a higher likelihood of positive returns, and thus words that co-occur with positive returns are assigned high values in O_+ and low values in O_-.

Our objective is to learn the model parameters, O_+, O_-, and p_i. In what follows, we detail three steps of the SSESTM procedure: 1) isolating the set of sentiment words, S, 2) estimating the topic parameters O_+ and O_-, and 3) estimating the article-level sentiment score p_i.

2.2 Screening for Sentiment-Charged Words

Sentiment-neutral words act as noise in our model, yet they are likely to dominate the data both in number of terms and in total counts. Estimating a topic model for the entire dictionary that accounts for the full joint distribution of sentiment-charged versus sentiment-neutral terms is at the very best a very challenging statistical problem, and at worst may suffer from severe inefficiency and high computational costs. Instead, our strategy is to isolate the subset of sentiment-charged words, and then estimate a topic model to this subset alone (leaving the neutral words unmodeled).

To accomplish this, we need an effective feature selection procedure to tease out words that carry sentiment information. We take a supervised approach that leverages the information in realized stock returns to screen for sentiment-charged words. Intuitively, if a word frequently co-occurs in articles that are accompanied by positive returns, that word is likely to convey positive sentiment.

Our screening procedure first calculates the frequency with which word j co-occurs with a positive return. This is measured as

$$f_j = \frac{\#\text{ articles including word } j \text{ AND having } \operatorname{sgn}(y) = 1}{\#\text{ articles including word } j} \tag{5}$$

for each $j = 1, ..., m$. Equivalently, f_j is the slope coefficient of a cross-article regression of $\operatorname{sgn}(y_i)$ on a dummy variable for whether word j appears in article i. This approach is known as marginal screening in the statistical literature (Fan and Lv, 2008). In comparison with the more complicated multivariate regression with sparse regularization, marginal screening is not only simple to use but also has a theoretical advantage when the signal to noise ratio is weak (Genovese et al., 2012; Ji and Jin, 2012).

Next, we set an upper threshold, α_+, and define all words having $f_j > \alpha_+$ as positive sentiment terms. Likewise, any word satisfying $f_j < \alpha_-$ for some lower threshold, α_-, is deemed a negative sentiment term. Finally, we select a third threshold, κ, on the count of articles including word j (i.e., the denominator of f_j, which we denote as k_j). Some sentiment words may appear infrequently in the data sample, in which case we have very noisy information about their relevance to sentiment. By restricting our analysis to words for which $k_j > \kappa$, we ensure minimal statistical accuracy of the frequency estimate, f_j. The thresholds $(\alpha_+, \alpha_-, \kappa)$ are hyper-parameters that can be tuned via cross-validation.[4]

Given $(\alpha_+, \alpha_-, \kappa)$, we construct the list of sentiment-charged terms that appropriately exceed these thresholds, which constitutes our estimate of the set S:[5]

[4] The definition in (5) is based on the number of articles, instead of the total number of word counts. In theory, one could threshold based on word count rather than article count, and this would have the same consistency property as our proposed method.

[5] In principle, we can combine our vocabulary with words identified in pre-existing sentiment dictionaries like Harvard-IV. To do this, one would expand $\widehat{S}$ to $\widetilde{S}$ according to:

$$\widetilde{S} = \widehat{S} \cup \{1 \le j \le m : \max\{\ell_j, 1 - \ell_j\} \ge \beta\}. \tag{6}$$

where $\ell \in [0,1]^m$ is a vector describing sentiment scores in the pre-existing dictionary, and β is a tunable threshold.

$$\widehat{S} = \{j : f_j \geq 1/2 + \alpha_+, \text{ or } f_j \leq 1/2 - \alpha_-\} \cap \{j : k_j \geq \kappa\}. \tag{7}$$

Algorithm 1 in Appendix A summarizes our screening procedure. Theorem C.2 of Appendix C establishes the procedure's "sure-screening" property, by which $\mathbb{P}(\widehat{S} = S)$ approaches one as the number of articles, n, and the number of words, m, jointly go to infinity (see, e.g. Fan and Lv, 2008).

2.3 Learning Sentiment Distributions

Once we have identified the relevant wordlist S, we arrive at the (now simplified) problem of fitting a two-topic model to the sentiment-charged counts. We can gather the two topic vectors in a matrix $O = [O_+, O_-]$, which determines the data generating process of the counts of sentiment-charged words in each article.

O captures information on both the frequency of words as well as their sentiment. It is helpful, in fact, to reorganize the topic vectors into a *vector of frequency*, F, and a *vector of tone*, T:

$$F = \frac{1}{2}(O_+ + O_-), \qquad T = \frac{1}{2}(O_+ - O_-), \tag{8}$$

If a word has a larger value in F, it appears more frequently overall. If a word has a larger value in T, its sentiment is more positive.

Classical topic models (Hofmann, 1999; Blei et al., 2003) amount to unsupervised reductions of the text, as these models do not assume availability of training labels for documents. Our setting differs from the classical setting because each Newswire is associated with a stock return. The returns contain information about the sentiment of articles, and hence returns serve as training labels. In a low signal-to-noise ratio environment, there are often large efficiency gains from exploiting document labels via supervised learning. We therefore take a supervised learning approach to estimate O (or, equivalently, to estimate F and T).

In our model, the parameter p_i is the article's sentiment score, as it describes how heavily the article tilts in favor of the positive word distribution. Suppose, for now, that we observe these sentiment scores for all articles in our sample. Let $\widetilde{d}_{i,[S]} = d_{i,[S]}/s_i$ denote the vector of word frequencies. Model (4) implies that

$$\mathbb{E}\widetilde{d}_{i,[S]} = \mathbb{E}\frac{d_{i,[S]}}{s_i} = p_i O_+ + (1 - p_i) O_-,$$

or, in matrix form,

$$\mathbb{E}\widetilde{D}' = OW, \qquad \text{where} \quad W = \begin{bmatrix} p_1 & \cdots & p_n \\ 1 - p_1 & \cdots & 1 - p_n \end{bmatrix}, \quad \text{and} \quad \widetilde{D} = [\widetilde{d}_1, \widetilde{d}_2, \ldots, \widetilde{d}_n]'.$$

Based on this fact, we propose a simple approach to estimate O via a regression of $\widetilde{D}$ on W. Note that we do not directly observe $\widetilde{D}$ (because S is unobserved) or W. We estimate $\widetilde{D}$ by plugging in $\widehat{S}$ from Algorithm 1. To estimate W, we use the standardized ranks of returns as sentiment scores for all articles in the training sample. More precisely, for each article i in the training sample $i = 1, ..., n$, we set

$$\widehat{p}_i = \frac{\text{rank of } y_i \text{ in } \{y_l\}_{l=1}^n}{n}. \tag{9}$$

and use these estimates to populate the matrix $\widehat{W}$. Intuitively, this estimator leverages the fact that the return y_i is a noisy signal for the sentiment of news in article i. This estimator, while obviously coarse, has a number of attractive features. First, it is simple to use and sufficient to achieve statistical guarantees for our algorithm under weak assumptions. Second, it is robust to outliers that riddle the return data.

Algorithm 2 in Appendix A summarizes our procedure for estimating $\widehat{O}$, and Theorem C.3 in Appendix C precisely characterizes the statistical accuracy of the algorithm. The algorithm consistently recovers the sentiment word frequency distribution, F. Its accuracy depends on the quality of the wordlist $\widehat{S}$ obtained from screening and the approximation quality of $\{\widehat{p}_i\}_{i=1}^n$ for $\{p_i\}_{i=1}^n$. The estimate of the tone vector, T, suffers a small bias that depends on the correlation between the true sentiment and the estimated sentiment, which takes the form

$$\rho = \frac{12}{n} \sum_{i=1}^{n} \Big(p_i - \frac{1}{2}\Big)\Big(\widehat{p}_i - \frac{1}{2}\Big). \tag{10}$$

Specifically, Theorem C.3 shows that the estimator $\widehat{T}$ converges to ρT. Therefore, when the estimation quality of $\widehat{p}$ is high, the bias is small. However, this scale bias has *no impact* on practical usage of the estimator. In practice, we are interested in the *relative* sentiment of words, not their *absolute* sentiment. The scalar multiple ρ washes out entirely when considering relative sentiment.

Given n articles realized from our topic model, with a vocabulary of size $|S|$ (i.e., the number of words in S), and an average article length (denoted $\bar{s}$), we show the convergence rate of the estimation errors of F and ρT are bounded by $\sqrt{|S|/(n\bar{s})}$, up to a logarithmic factor. In our empirical study, the identified sentiment dictionary contains approximately 100 to 200 words, yet their total count in one article is typically below 20. So we are primarily interested in the "short article" case, that is, $\bar{s}/|S| \leq C$ for some constant C, as opposed to the "long article" case, in which $\bar{s}/|S| \to \infty$. As shown in Ke and Wang (2017), the classical unsupervised approach converges at a slower rate than ours in the case of short articles. The statistical efficiency gain of supervised learning in the short article setting is the central consideration behind our choice of a supervised topic modeling approach.

2.4 Scoring New Articles

The preceding steps construct estimators $\widehat{S}$ and $\widehat{O}$. We now discuss how to estimate the sentiment p_i for a new article i that is not included from the training sample.

Let d_i be the article's count vector and let s_i be its total count of sentiment-charged words. Given our model (4),

$$d_{i,[S]} \sim \text{Multinomial}\Big(s_i,\ p_i O_+ + (1 - p_i) O_-\Big),$$

and given $\widehat{S}$ and $\widehat{O}$, we can estimate p_i using maximum likelihood estimation (MLE). While alternative estimators, such as linear regression, are also consistent, we use MLE for its statistical efficiency.

We add a penalty term, $\lambda \log(p_i(1-p_i))$, in the likelihood function, which is described explicitly in (A.3) of Algorithm 3. The role of the penalty is to help cope with the limited number of observations and the low signal-to-noise ratio inherent to return prediction. Imposing the penalty shrinks the estimate towards a neutral sentiment score of 1/2, where the amount of shrinkage depends on the magnitude of λ.[6] This penalized likelihood approach is equivalent to imposing a Beta distribution prior on the sentiment score. Most articles have neutral sentiment, and the beta prior ensures that this is reflected in the model estimates.

Theorem C.4 in Appendix C provides a statistical guarantee for our scoring procedure. Not surprisingly given our earlier discussion, the estimator is inconsistent with respect to p_i, and instead converges to $\frac{1}{2} + \frac{1}{\rho}\left(p_i - \frac{1}{2}\right)$ instead. The inflation factor of $1/\rho$ arises from the bias in estimating T. Our penalization is expressly intended to help deflate these estimates. And, as emphasized earlier, this has no impact on the relative sentiment scores of new articles (or on our application to portfolio choice). In terms of the convergence rate, besides the estimation error accumulated from the previous two steps, an additional error of magnitude $1/\sqrt{s}$ appears. Intuitively, if the article contains very few sentiment words, its sentiment score will not be accurately recovered. And again, in such circumstances, penalization serves to improve efficiency.

3 Empirical Analysis

In this section, we apply our text-mining framework to the problem of return prediction for investment portfolio construction. This application serves two purposes. First, it offers an empirical demonstration of the predictive content in our sentiment measure. Second, it translates the extent of predictability from statistical terms such as predictive R^2, into the more meaningful economic terms of growth rates in an investor's savings portfolio attributable to exploitation of text-based information.

To develop hypotheses, it is useful to consider the potential sources of time series return predictability from news text. A natural null hypothesis for any return prediction analysis is the efficient markets hypothesis (Fama, 1970). Market efficiency predicts that the expected return is dominated by unforecastable news, as this news is rapidly (in its starkest form, immediately) and fully incorporated in prices. The maintained hypothesis of our research is that information in news text is not fully absorbed by market prices instantaneously, for reasons such as limits-to-arbitrage and rationally limited attention. As a result, information contained in news text is predictive of future asset price paths, at least over short horizons. While this alternative hypothesis is by now uncontroversial, it is hard to overstate its importance, as we have much to learn about the mechanisms through which information enters prices and the frictions that impede these mechanisms. Our prediction analysis adds new evidence to the empirical literature investigating the alternative hypothesis. In particular,

[6] The single penalty parameter λ is common across articles. This implies that the relative ranks of article sentiment are not influenced by penalization, which is the key information input into the trading strategy in our empirical analysis.

Table 1: Summary Statistics

Filter	Remaining Sample Size	Observations Removed
Total Number of Dow Jones Newswire Articles	$31,492,473$	
Combine chained articles	$22,471,222$	$9,021,251$
Remove articles with no stocks tagged	$14,044,812$	$8,426,410$
Remove articles with more than one stocks tagged	$10,364,189$	$3,680,623$
Number of articles whose tagged stocks have three consecutive daily returns from CRSP between Jan 1989 and Dec 2012	6,540,036	
Number of articles whose tagged stocks have open-to-open returns from CRSP since Feb 2004	6,790,592	
Number of articles whose tagged stocks have high-frequency returns from TAQ since Feb 2004	6,708,077	

Note: In this table, we report the impact of each filter we apply on the number of articles in our sample. The sample period ranges from January 1, 1989 to July 31, 2017. The CRSP three-day returns are only used in training and validation steps, so we apply the CRSP filter only for articles dated from January 1, 1989 to December 31, 2012. The open-to-open returns and intraday returns are used in out-of-sample periods from February 1, 2004 to July 31, 2017.

we bring to bear information from a rich and hitherto unstudied news text data set. Our methodological contribution is a new toolkit for that makes it feasible to conduct a coherent statistical analysis of such complex and unstructured data. An ideal (and hopefully realizable) outcome of our research agenda is to understand how news influences investors' belief formation and then enters prices.

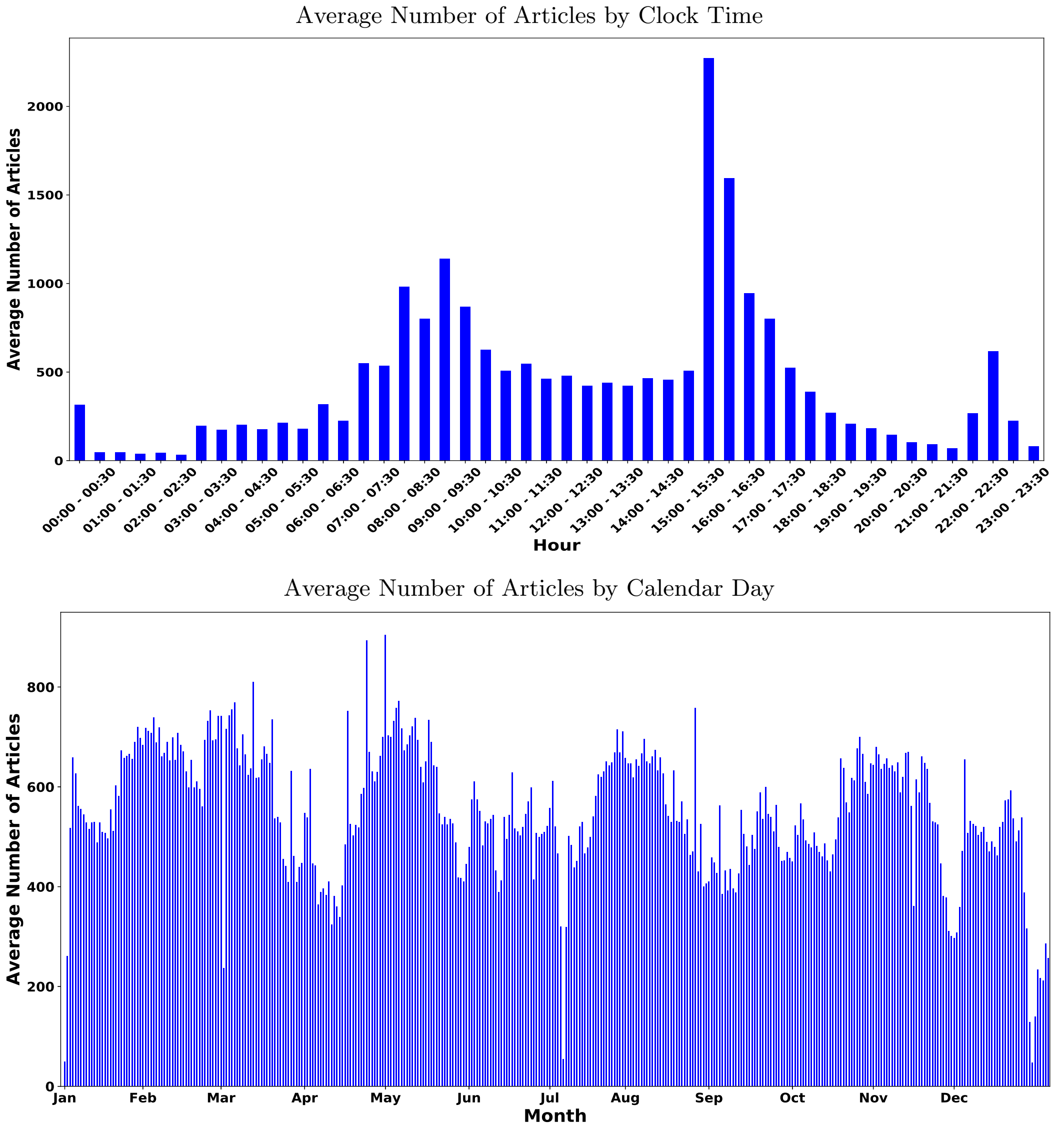
3.1 Data and Pre-processing

Our text data set is the *Dow Jones Newswires Machine Text Feed and Archive* database. It contains real-time news feeds from January 1, 1989 to July 31, 2017, amounting to 22,471,222 unique articles (after combining "chained" articles). Approximately 62.5% news articles are assigned one or more firm tags describing the primary firms to which the article pertains. To most closely align the data with our model structure, we remove articles with more than one firm tag, or 16.4% articles, arriving at a sample of 10,364,189 articles. We track the date, exact timestamp, tagged firm ticker, headline, and article body of each article.

Using ticker tags, we match each article with tagged firm's market capitalization and adjusted daily close-to-close returns from CRSP. We do not know, a priori, the timing by which potential new information in a Newswire article gets impounded in prices. If prices adjust slowly, then it makes sense to align articles with contemporaneous or even future returns. Newswires are a highly visible information source for market participants, so presumably any delay in price response would be short-lived, perhaps on the order of a day. Or, it could be the case that Newswires are a restatement of recently revealed information, in which case news is best aligned with prior day returns.

Without better guidance on timing choice, we train the model by merging articles published

Figure 2: Average Article Counts



Average Number of Articles by Clock Time

Average Number of Articles by Calendar Day

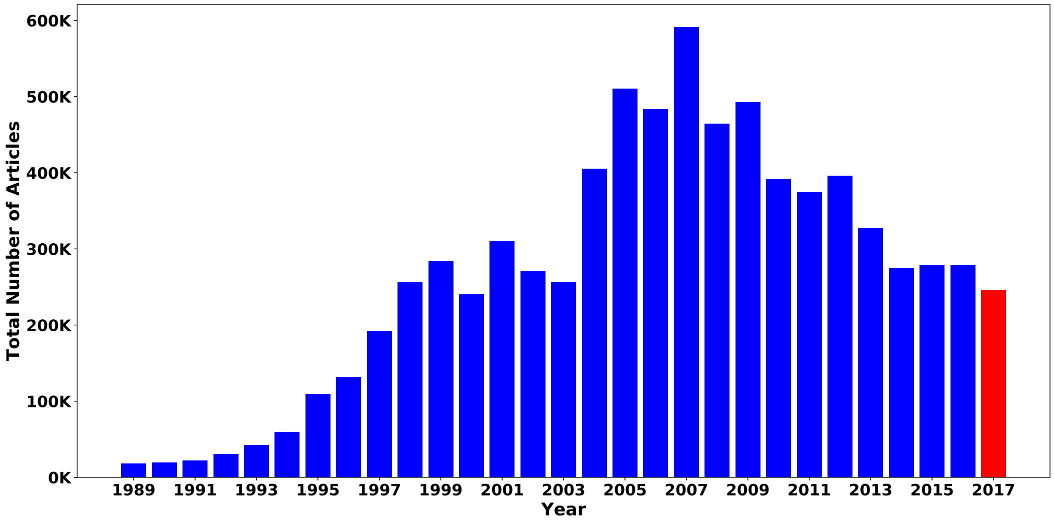
Note: The top figure plots the average numbers of articles per half an hour (24 hour EST time) from January 1, 1989 to July 31, 2017. The bottom figure plots the average numbers of articles per calendar day. Averages are taken over the full sample from January 1, 1987 to July 31, 2017.

between 4pm of day $t-1$ and 4pm of day t with the tagged firm's three-day return from close on day $t-2$ to close on day $t+1$.[7] Note that this timing is for sentiment training purposes only. In order to devise a trading strategy, for example, it is critical to align sentiment estimates for an article *only* with future realized returns (as we discuss further below).

For some of our analyses we study the association between news text and intradaily returns. For this purpose, we merge articles with transaction prices from the NYSE Trade and Quote (TAQ)

[7] For news that occur on holidays or weekends, we use the next available trading day as the current day t and the last trading day before the news as day $t-1$.

Figure 3: Annual Time Series of the Total Number of Articles



Note: This figure plots the annual time series of the total number of articles from January 1987 to July 2017. We only provide an estimate for 2017 (highlighted in red), by annualizing the total number of articles of the few months we observe, since we do not have a whole year's data for this year.

database. Open-to-open and intraday returns are only used in our out-of-sample analysis from February 2004 to July 2017. We start the out-of-sample testing period from February 2004 because, starting in January 17, 2004, the Newswire data is streamlined and comes exclusively from one data source. Prior to that, Newswires data are derived from multiple news sources, which among other things can lead to redundant coverage of the same event. Although it does not affect in-sample training and validation, this could have an adverse impact on our out-of-sample analysis that is best suited for "fresh" news. In summary, Table 1 lists step-by-step details for our sample filters.

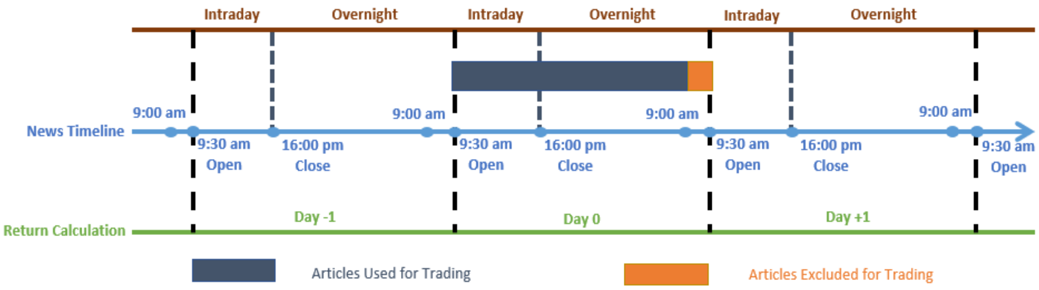
The top panel of Figure 2 plots the average number of articles in each half hour throughout a day. News articles arrive more frequently prior to the market open and close. The bottom panel plots the average number of articles per day over a year. It shows leap-year and holiday effects, as well as quarterly earnings season effects corresponding to a rise in article counts around February, May, August, and November. Figure 3 plots the total number of news articles per year in our sample. There is a steady increase in the number of news articles until around 2007. Some news volume patterns reflect structural changes in data sources. According to the *Dow Jones Newswires* user guide, there were three historical merges of news sources which occurred on October 31, 1996, November 5, 2001, and January 16, 2004, respectively.

Next, we follow common steps from the natural language processing literature to clean and structure news articles.[8] The first step is normalization, including 1) changing all words in the article to lower case letters; 2) expanding contractions such as "haven't" to "have not"; and 3) deleting numbers, punctuations, special symbols, and non-English words.[9] The second step is stemming and

[8] We use the natural language toolkit (NLTK) in Python to preprocess the data.

[9] The list of English words is available from item 61 on http://www.nltk.org/nltk_data/.

Figure 4: News Timeline



Note: This figure describes the news timeline and our trading activities. We exclude news from 9:00 am to 9:30 am EST from trading (our testing exercise), although these news are still used for training and validation purposes. For news that occur on day 0, we build positions at the market opening on day 1, and rebalance at the next market opening, holding the positions of the portfolio within the day. We call this portfolio day+1 portfolio. Similarly, we can define day 0 and day-1, day± 2, $\dots$, day± 10 portfolios.

lemmatizing, which group together the different forms of a word to analyze them as a single root word, e.g., “disappointment” to “disappoint,” “likes” to “like,” and so forth.[10] The third step is tokenization, which splits each article into a list of words. The fourth step removes common stop words such as “and”, “the”, “is”, and “are.”[11] Finally, we translate each article into a vector of word counts, which constitute its so-called “bag of words” representation.

We also obtain a list of 2,337 negative words (Fin-Neg) and 353 positive words (Fin-Pos) from the Loughran-McDonald (LM) Sentiment Word Lists for comparison purposes.[12] LM show that the Harvard-IV misclassifies words when gauging tone in financial applications, and propose their own dictionary for use in business and financial contexts.

3.2 Return Predictions

We train the model using rolling window estimation. The rolling estimation window consists of a 15 year interval, the first 10 years of which are used for training with the last five years used for validation/tuning. We then use the subsequent one-year window for out-of-sample testing. At the end of the testing year, we roll the entire analysis forward by a year and re-train. We iterate this procedure until we exhaust the full sample, which amounts to estimating and validating the model 14 times.

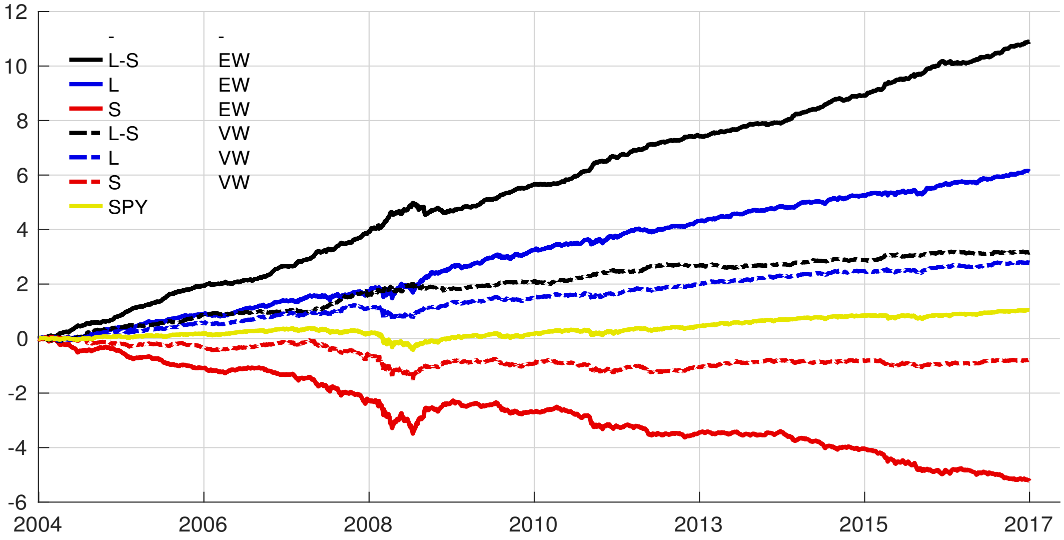
In each training sample, we estimate a collection of SSESTM models corresponding to a grid

[10] The lemmatization procedure uses the WordNet as a reference database: https://wordnet.princeton.edu/. The stemming procedure uses the package “porter2stemmer” on https://pypi.org/project/porter2stemmer/. Frequently, the stem of an English word is not itself an English word; for example, the stem of “accretive” and “accretion” is “accret.” In such cases, we replace the root with the most frequent variant of that stem in our sample (e.g., “accretion”) among all words sharing the same stem, which aids interpretability of estimation output.

[11] We use the list of stopwords available from item 70 on http://www.nltk.org/nltk_data/.

[12] The Loughran-McDonald word lists also include 285 words in Fin-Unc, 731 words in Fin-Lit, 19 strong modal words and 27 weak words. We only present results based on Fin-Neg and Fin-Pos. Other dictionaries are less relevant to sentiment.

Figure 5: One-day-ahead Performance Comparison of SSESTM



Note: This figure compares the cumulative log returns of portfolios sorted on out-of-sample sentiment scores. The black, blue, and red colors represent the long-short (L-S), long (L), and short (S) portfolios, respectively. The solid and dashed lines represent equal-weighted (EW) and value-weighted (VW) portfolios, respectively. The yellow solid line is the S&P 500 return (SPY). The legend provides a table of annual Sharpe ratios (SR) and average daily returns (AvgRet) in basis points.

of tuning parameters.[13] We use all estimated models to score each news article in the validation sample, and select the constellation of tuning parameter values that minimizes a loss function in the validation sample. Our loss function is the ℓ^1-norm of the differences between estimated article sentiment scores and the corresponding standardized return ranks for all events in the validation sample.

3.3 Daily Predictions

Figure 5 reports the cumulative one-day trading strategy returns (calculated from open-to-open) based on out-of-sample SSESTM sentiment forecasts. We report the long (L) and short (S) sides separately, as well as the overall long-short (L-S) strategy performance. We also contrast performance of equal-weighted (EW) and value-weighted (VW) versions of the strategy. Table 2 reports the corresponding summary statistics of these portfolios in detail.

In the out-of-sample test period, we estimate the sentiment scores of articles using the optimally tuned model determined from the validation sample. In the case a stock is mentioned in multiple news articles on the same day, we forecast the next-day return using the average sentiment score over the coincident articles. For firms that have no news, we assign a neutral score (0.5).

[13] There are four tuning parameters in our model, including $(\alpha_+, \alpha_-, \kappa, \lambda)$. We consider three choices for α_+ and α_-, which are always set such that the number of words in each group (positive and negative) is either 25, 50, or 100. We consider five choices of κ ($86\%, 88\%, 90\%, 92\%$, and 94% quantiles of the count distribution each year), and three choices of λ (1, 5, and 10).

Table 2: Performance of Daily News Sentiment Portfolios

	Sharpe		Average	FF3		FF5		FF5+MOM	
Formation	Ratio	Turnover	Return	α	R^2	α	R^2	α	R^2
EW L-S	4.29	94.6%	33	33	1.8%	32	3.0%	32	4.3%
EW L	2.12	95.8%	19	16	40.0%	16	40.3%	17	41.1%
EW S	1.21	93.4%	14	17	33.2%	16	34.2%	16	36.3%
VW L-S	1.33	91.4%	10	10	7.9%	10	9.3%	10	10.0%
VW L	1.06	93.2%	9	7	30.7%	7	30.8%	7	30.8%
VW S	0.04	89.7%	1	4	31.8%	3	32.4%	3	32.9%

Note: The table report the performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios and their long (L) and short (S) legs. The performance measures include (annualized) annual Sharpe ratio, annualized expected returns, risk-adjusted alphas, and R^2s with respect the Fama-French three-factor model ("FF3"), the Fama-French five-factor model ("FF5'), and the Fama-French five-factor model augmented to include the momentum factor ("FF5+MOM"). We also report the strategy's daily turnover, defined as $\frac{1}{2T}\sum_{t=1}^{T}\left(\sum_i |w_{i,t+1} - w_{i,t}(1+r_{i,t+1})|\right)$, where $w_{i,t}$ is the weight of stock i in the portfolio at time t.

To evaluate out-of-sample predictive performance in economic terms, we design a trading strategy that leverages sentiment estimates for prediction. Our trading strategy is very simple. It is a zero-net-investment portfolio that each day buys the 50 stocks with the most positive sentiment scores and shorts the 50 stocks with the most negative sentiment scores.[14]

We consider both equal-weighted and value-weighted schemes when forming the long and short sides of the strategy. Equal weighting is a simple and robust means of assessing predictive power of sentiment throughout the firm size spectrum, and is anecdotally closer to the way that hedge funds use news text for portfolio construction. Value weighting heavily overweights large stocks, which may be justifiable for economic reasons (assigning more weight to more productive firms) and for practical trade implementation reasons (such as limiting transaction costs).

We form portfolios every day, and hold them for anywhere from a few hours up to ten days. We are careful to form portfolios only at the market open each day for two reasons. First, overnight news can be challenging to act on prior to the morning open as this is the earliest time most traders can access the market. Second, with the exception of funds that specialize in high-frequency trading, funds are unlikely to change their positions continuously in response to intraday news because of their investment styles and investment process constraints. Finally, following a similar choice of Tetlock et al. (2008), we exclude articles published between 9:00am and 9:30am EST. By imposing that trade occurs at the market open and with at least a half-hour delay, we hope to better match realistic considerations like allowing funds time to calculate their positions in response to news and allowing them to trade when liquidity tends to be highest. Figure 4 summarizes the news and trading timing of our approach.

Three basic facts emerge from the one-day forecast evaluation. First, equal-weighted portfolios substantially outperform their value-weighted counterparts. The long-short strategy with equal weights earns an annualized Sharpe ratio of 4.29, versus 1.33 in the value-weighted case. This in-

[14] On a few days there are fewer than 50 firms with non-neutral scores, in which case we trade fewer than 100 stocks but otherwise maintain the zero-cost nature of the portfolio.

dicates that news article sentiment is a stronger predictor of future returns to small stocks, all else equal. There are a number of potential economic explanations for this fact. It may arise, for example, due to i) the fact that small stocks receive less investor attention and thus respond slower to news, ii) that the underlying fundamentals of small stocks are more uncertain and opaque and thus require more effort to process news into actionable price assessments, or iii) that small stocks are less liquid and thereby require a longer time for trading to occur to incorporate information into prices.

Second, the long side of the trade outperforms the short side, with a Sharpe ratio 2.12 versus 1.21 (in the equal-weighted case). This fact is in part due to the fact that the long side naturally earns the market equity risk premium while the short side pays it. A further potential explanation is that investors face short sales constraints. Our findings are consistent with Santosh (2016), who shows less reaction to negative earnings announcements, and Reed (2007), who finds slower price reaction among firms with bad earnings news particularly for those stocks that are more difficult to short.

Third, the SSESTM sentiment trading strategies have little exposure to standard aggregate risk factors. The individual long and short legs of the trade have at most a 41% daily R^2 when regressed on Fama-French factors, while the long-short spread portfolio R^2 is at most 10%. In all cases, the average return of the strategy is almost entirely alpha. Note that, by construction, the daily turnover of the portfolio is large. If we completely liquidated the portfolio at the end of each day, we would have a turnover of 100% per day. Actual turnover is slightly lower, on the order of 94% for equal-weighted implementation and 90% for value-weighted, indicating a small amount of persistence in positions. In the value-weighted case, for example, roughly one in ten stock trades is kept on for two days—these are instances in which news of the same sentiment for the same firm arrives in successive days. Finally, Figure 5 shows that the long-short strategy avoids major drawdowns, and indeed appreciates during the financial crisis while SPY sells off.

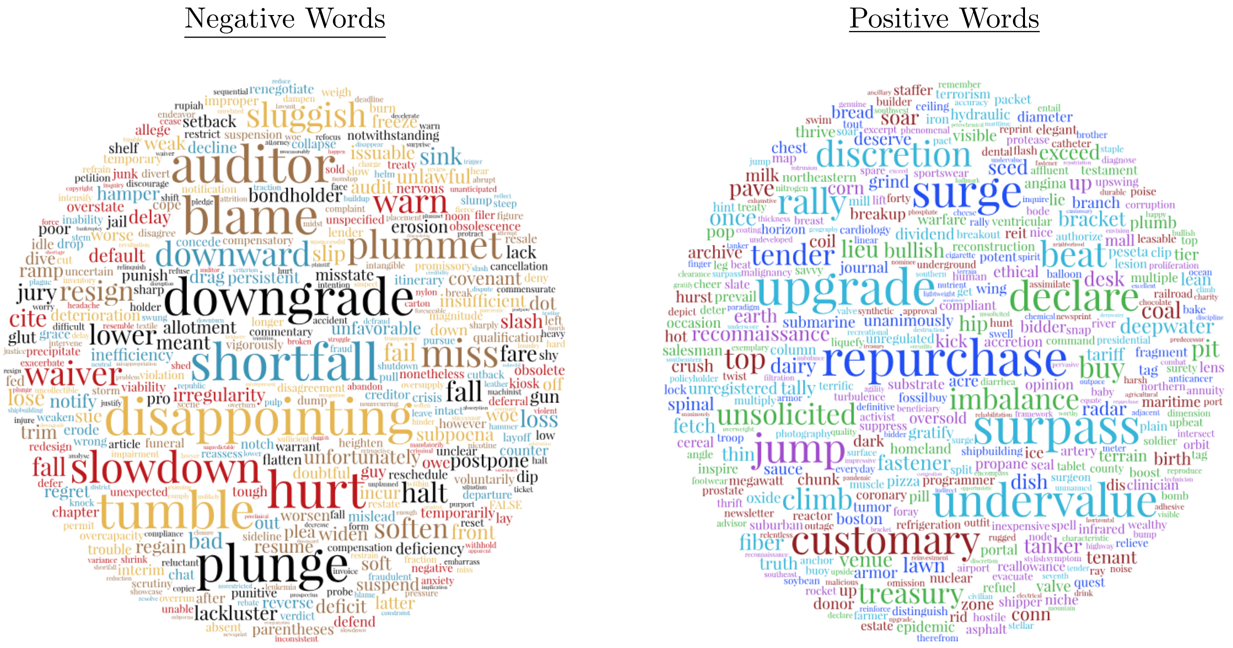
3.4 Most Impactful Words

Figure 6 reports the list of sentiment-charged words estimated from our model. These are the words that most strongly correlate with realized price fluctuations and thus surpass the correlation screening threshold. Because we re-estimate the model in each of our 14 training samples, the sentiment word lists can change throughout our analysis. To illustrate the most impactful sentiment words in our analysis, the word cloud font is drawn proportional to the number of training samples (out of 14) that the word appears in the corresponding sentiment list. Table A.2 in Appendix F provides additional detail on selected words, reporting the top 50 positive and negative sentiment words over all training samples. It reports the number of training samples each word is selected and its average sentiment score in the selected samples.

The estimated wordlists are remarkably stable over time. Of the top 50 positive sentiment words over all periods, 25 are selected into the positively charged set in at least 9 of the 14 training samples. For the 50 most negative sentiment words, 25 are selected in at least 7 out of 14 samples. The following nine negative words are selected in *every* training sample:

shortfall, downgrade, disappointing, tumble, blame, hurt, auditor, plunge, slowdown,

Figure 6: Sentiment-charged Words



Negative Words

Positive Words

Note: This figure reports the list of words in the sentiment-charged set S. Font size of a word is proportional to the number of training samples (out of 14) that the word appears in the corresponding sentiment list.

and the following words are selected into the positive word ten or more training samples:

repurchase, surpass, upgrade, undervalue, surge, customary, jump, declare, rally, discretion, beat.

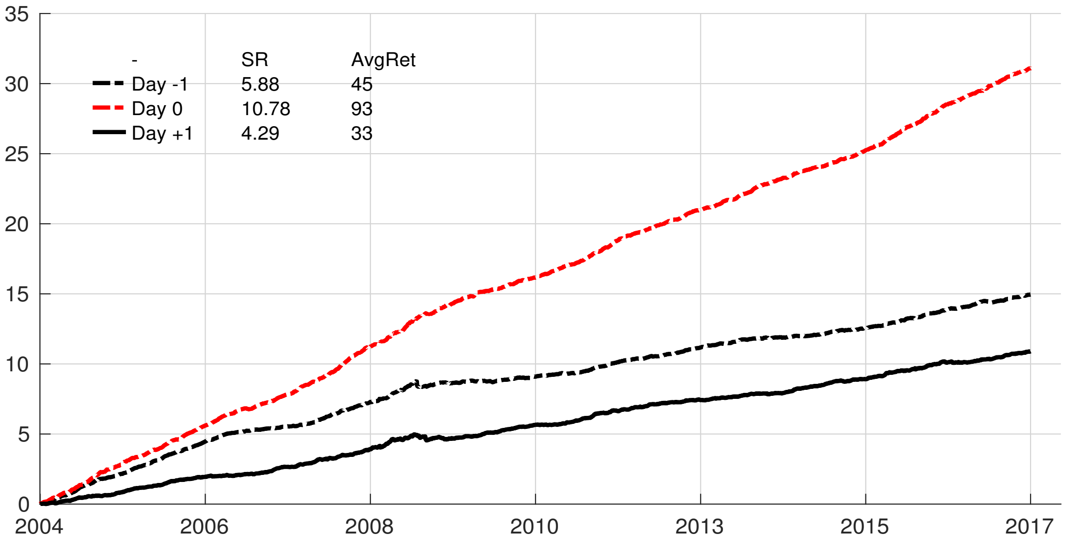
There are interesting distinctions vis-a-vis extant sentiment dictionaries. For example, in comparison to our estimated list of the eleven most impactful positive words listed above, only one (*surpass*) appears in the LM positive dictionary, and only four (*surpass, upgrade, surge, discretion*) appear in Harvard-IV. Likewise, four of our most impactful negative terms (*tumble, blame, auditor, plunge*) do not appear in the LM negative dictionary and six are absent from Harvard-IV. Thus, in addition to the fact that our word lists are accompanied by term specific sentiment weights (contrasting with the implicit equal weights in extant dictionaries), many of the words that we estimate to be most important for understanding realized returns are entirely omitted from pre-existing dictionaries.

3.5 Fresh News and Stale News

The analysis in Figure 5 and Table 2 focuses on relating news sentiment on day t to returns on day $t+1$. In this section, we investigate the timing of price responses to news sentiment more precisely.

Recall that in our training sample, we estimate SSESTM from the three-day return beginning the day before an article is published and ending the day after. In Figure 7, we separately investigate the association between news sentiment on day t and returns on day $t-1$ (from open $t-1$ to open t), day t, and day $t+1$. We report this association in the economic terms of trading strategy

Figure 7: Price Response On Days −1, 0, and 1



Note: This figure compares the cumulative log returns of long-short portfolios sorted on out-of-sample sentiment scores. The Day −1 strategy (dashed black line) shows the association between news and returns one day prior to the news; the Day 0 strategy (dashed red line) shows the association between news and returns on the same day; and the Day +1 strategy (solid black line) shows the association between news and returns one day later. The Day −1 and Day 0 strategy performance is out-of-sample in that the model is trained on a sample that entirely precedes portfolio formation, but these are not implementable strategies because the timing of the news article would not necessarily allow a trader to take such positions in real time. They are instead interpreted as out-of-sample correlations between article sentiment and realized returns in economic return units. The Day +1 strategy corresponds to the implementable trading strategy shown in Figure 5. All strategies are equal-weighted. The legend reports annualized Sharpe ratios (SR) and average daily returns (AvgRet) in basis points.

performance. The association between sentiment and the $t+1$ return is identical to that in Figure 5, and is rightly interpreted as performance of an implementable (out-of-sample) trading strategy. For the association with returns on days $t-1$ and t, the interpretation is different. These are *not* implementable strategies because the timing of the news article would not necessarily allow a trader to take a position and exploit the return at time t (and certainly not at $t-1$). They are instead interpreted as out-of-sample correlations between article sentiment and realized returns, converted into economic return units. They are out-of-sample because the fitted article sentiment score, $\hat{p}_i$, is based on a model estimated from an entirely different data set (that distinctly pre-dates the arrival of article i and returns $r_{i,t-1}$, $r_{i,t}$, and $r_{i,t+1}$). Table 3 reports summary statistics for these portfolios, including their annualized Sharpe ratios, average returns, alphas, and turnover. For this analysis, we specialize to equal-weighted value-weighted portfolios.

The Day −1 strategy (dashed black line) shows the association between news article sentiment, and the stock return one day prior to the news. This strategy thus quantifies the extent to which our sentiment score picks up on stale news. On average, prices respond strongly ahead of news in our sample, as indicated by the infeasible annualized Sharpe ratio of 5.88. Thus we see that much of the daily news flow echoes previously reported news or is a new report of information already known to market participants.

Table 3: Price Response On Days −1, 0, and +1

	Sharpe		Average	FF3		FF5		FF5+MOM	
Formation	Ratio	Turnover	Return	α	R^2	α	R^2	α	R^2
				Day −1					
L-S	5.88	94.5%	45	45	0.1%	44	0.5%	44	0.6%
L	2.30	95.9%	20	20	0.8%	21	1.1%	21	1.1%
S	2.08	93.2%	25	24	0.5%	24	1.2%	24	1.2%
				Day 0					
L-S	10.78	94.6%	93	93	0.4%	93	0.5%	92	0.8%
L	5.34	96.0%	50	48	7.0%	49	7.8%	49	8.1%
S	3.56	93.3%	43	45	6.0%	44	7.0%	43	7.5%
				Day +1					
L-S	4.29	94.6%	33	33	1.8%	32	3.0%	32	4.3%
L	2.12	95.8%	19	16	40.0%	16	40.3%	17	41.1%
S	1.21	93.4%	14	17	33.2%	16	34.2%	16	36.3%
				Day −1 to +1					
L-S	12.38	94.6%	170	170	1.0%	169	2.3%	169	2.8%
L	5.67	95.9%	89	86	22.3%	86	23.2%	87	24.1%
S	3.83	93.3%	81	85	16.7%	82	18.7%	82	20.1%

Note: The table repeats the analysis of Table 2 for the equal-weighted long-short (L-S) portfolios plotted in Figure 7, as well as their long (L) and short (S) legs. Sharpe ratios are annualized, while returns and alphas are in basis points per day.

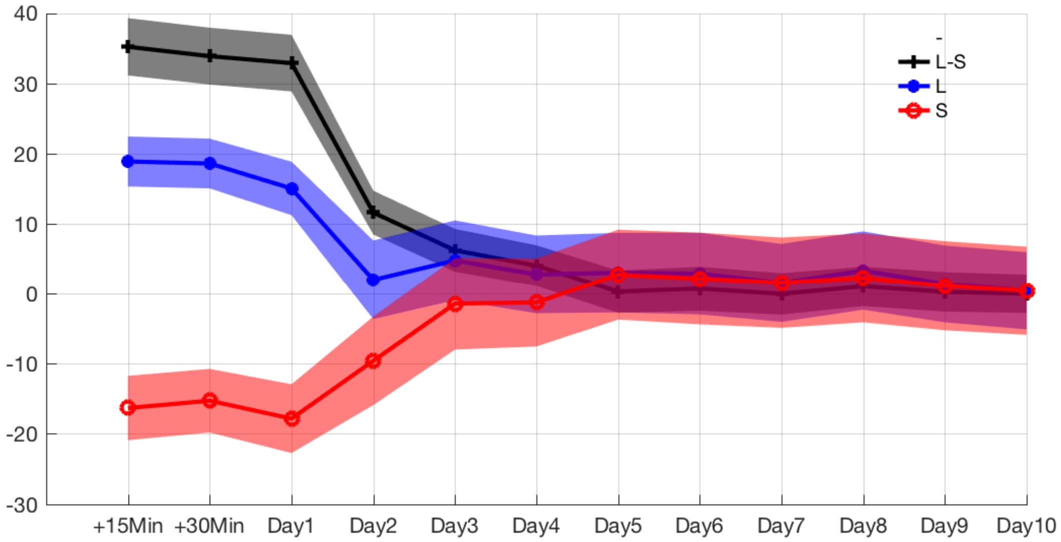
The Day 0 strategy (dashed red line) shows the association between news and returns on the same day. This strategy assesses the extent to which our sentiment score captures fresh news that has not previously been incorporated into prices. The Day 0 strategy provides the clearest out-of-sample validation that our sentiment score accurately summarizes fresh, value-relevant information content of news text. Indeed, we see that the strongest price responses occur on the same day that the news arrives, reflected in the infeasible annualized Sharpe ratio of 10.78.

The Day +1 strategy (solid black line) shows the association between news and returns one day later, and thus quantifies the extent to which information in our sentiment score is impounded into prices with a delay. This corresponds exactly to the implementable trading strategy shown in Figure 5. The excess performance of this strategy, summarized in terms of an annualized Sharpe ratio of 4.29 (and shown to be all alpha in Table 2), supports the maintained hypothesis.

3.6 Speed of Information Assimilation

We refine our analysis of the maintained hypothesis with a more granular investigation of the timing of price responses to news sentiment. In particular, we analyze trading strategies that trade in response to news sentiment with various time delays. We consider very rapid price responses via intra-day high frequency trading that takes a position either 15 or 30 minutes after the article's time stamp, and holds positions until the next day's open. We also consider daily frequencies by studying performance of daily open-to-open returns initiated from one to 10 days following the announcement,

Figure 8: Speed of News Assimilation



Note: This figure compares average one-day holding period returns to the news sentiment trading strategy as a function of when the trade is initiated. We consider intra-day high frequency trading that takes place either 15 or 30 minutes after the article's time stamp and is held for one day (denoted +15min and +30min, respectively), and daily open-to-open returns initiated from one to 10 days following the announcement. We report equal-weighted portfolio average returns (in basis points per day) in excess of an equal-weighted version of the S&P 500 index, with 95% confidence intervals given by the shaded regions. We consider the long-short (L-S) portfolio as well as the long (L) and short (S) legs separately.

in each case holding the position for one day.

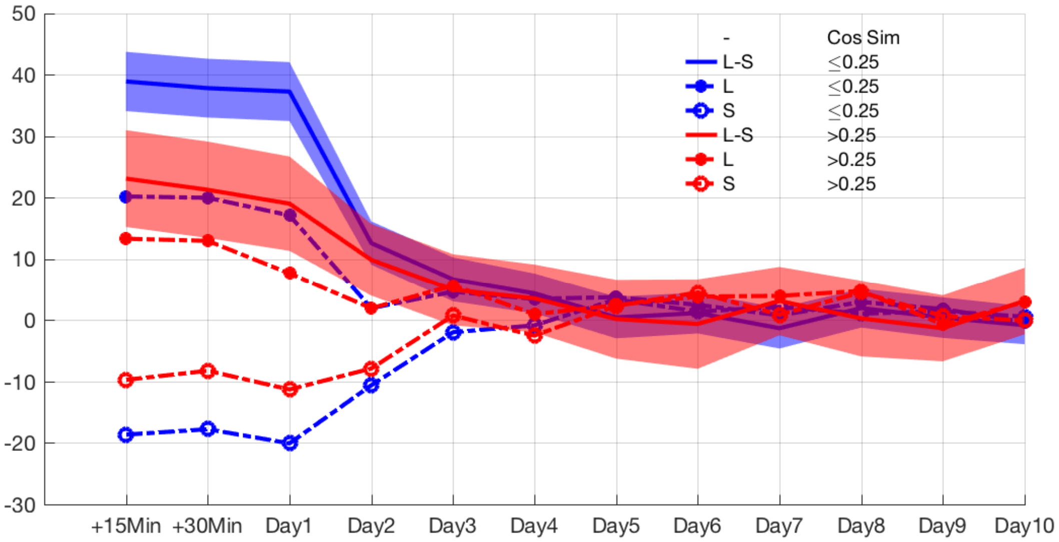
Figure 8 reports average returns in basis points per day with 95% confidence intervals given by the shaded regions. It shows the long-short portfolio as well as the long and short legs separately. The results show that, for the long-short strategy, sentiment information is essentially fully incorporated into prices by the start of Day +3. For the individual sides of the trade, the long leg appears to achieve full price incorporation within two days, while the short leg takes one extra day.

The evidence in Section 3.5 indicates that a substantial fraction news is "old news" and already impounded in prices by the time an article is published, and the assimilation analysis of Figure 8 does not distinguish between fresh versus stale news. In order to investigate the difference in price response to fresh versus stale news, we construct a measure of article novelty as follows.

For each article for firm i on day t, we calculate its cosine similarity with all articles about firm i on the five trading days prior to t (denoted by the set $\chi_{i,t}$). Novelty of recent news is judged based on its most similar preceding article, thus we define article novelty as

$$\text{Novelty}_{i,t} = 1 - \max_{j \in \chi_{i,t}} \left(\frac{d_{i,t} \cdot d_j}{\|d_{i,t}\| \, \|d_j\|} \right).$$

Figure 9: Speed of News Assimilation (Fresh Versus Stale News)



Note: See Figure 8. This figure divides stock-level news events based on maximum cosine similarity with the stock's prior news.

Figure 9 separates price assimilation of news based on article novelty. We partition news into two groups. "Fresh" news has a maximum cosine similarity of 0.25 with previous articles, while "stale" news has a maximum cosine similarity greater than 0.25.[15] It shows that the one-day price response (from fifteen minutes after news arrival to the open the following day) of the long-short portfolio formed on fresh news (solid blue line) is 39 basis points, nearly doubling the response of 23 basis points to stale news (solid red line). Furthermore, it takes four days for fresh news to be fully incorporated in prices (i.e., at day five average return is statistically indistinguishable from zero), or twice as long as the two days it takes for prices to complete their response to stale news.

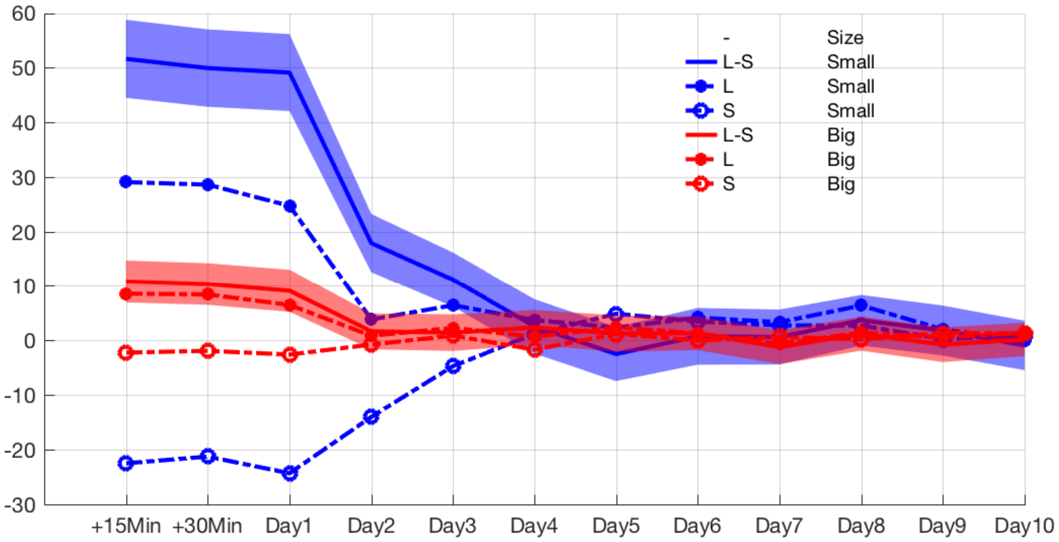
3.7 Stock Heterogeneity Analysis: Stock Size and Volatility

Figure 9 investigates differential price responses to different types of news. In this section, we investigate differences in price assimilation with respect to heterogeneity among stocks.

The first dimension of stock heterogeneity that we analyze is market capitalization. Larger stocks represent a larger share of the representative investor's wealth and command larger fraction of investors' attention or information acquisition effort (e.g., Wilson, 1975; Veldkamp, 2006). In Figure 10, we analyze the differences in price adjustment based on firm size by sorting stocks into big and small groups (based on NYSE median market capitalization each period). Prices of large

[15] The average similarity measure in our sample is approximately 0.25. The conclusions from Figure 9 are generally insensitive the choice of cutoff.

Figure 10: Speed of News Assimilation (Big Versus Small Stocks)



Note: See Figure 8. This figure divides stock-level news events based on stocks' market capitalization. The big/small breakpoint is defined as the NYSE median market capitalization each period.

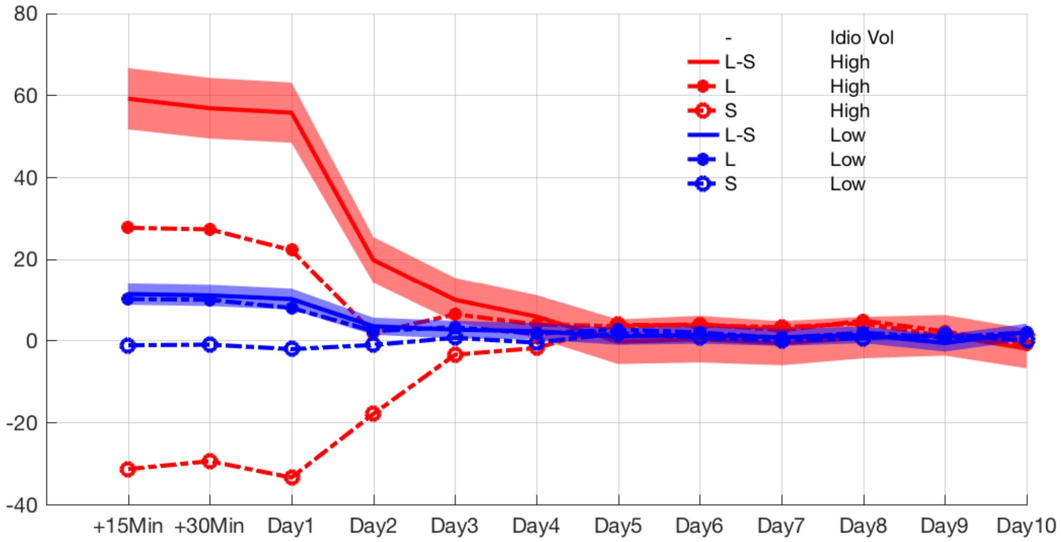
stocks respond by 11 basis points in the first day after news arrival, and their price response is complete after one day (the day two effect is insignificantly different from zero). The price response of small stocks is 52 basis points in the first fifteen minutes, nearly five times larger, and it take three days for their news to be fully incorporated into prices.

The second dimension of heterogeneity that we investigate is stock volatility. On one hand, it is a limit to arbitrage, as higher volatility dissuades traders from taking a position based on their information, all else equal. At the same time, higher stock volatility represents more uncertainty about asset outcomes. With more uncertainty, there are potentially larger profits to be earned by investors with superior information, which incentivizes informed investors to allocate more attention to volatile stocks all else equal. But higher uncertainty may also reflect that news about the stock is more difficult to interpret, manifesting in slower incorporation into prices. The direction of this effect on price assimilation is ambiguous.

Figure 11 shows the comparative price response of high versus low volatility firms.[16] The price response to SSESTM sentiment in the first 15 minutes following news arrival is 11 basis points for low volatility firms, but 52 basis points for high volatility firms. And while news about low volatility firms is fully impounded in prices after one day of trading, it takes three days for news to be fully

[16] Specifically, we calculate idiosyncratic volatility from residuals of a market model using the preceding 250 daily return observations. We then estimate the conditional idiosyncratic volatility with via exponential smoothing according to the formula $\sigma_t = \sum_{i=0}^{\infty}(1-\delta)\delta^i u_{t-1-i}^2$ where u is the market model residual and δ is chosen so that the exponentially-weighted moving average has a center of mass $(\delta/(1-\delta))$ of 60 days .

Figure 11: Speed of News Assimilation (High Versus Low Volatility Stocks)



Note: See Figure 8. This figure divides stock-level news events based on stocks' idiosyncratic volatility. The high/low volatility breakpoint is defined as the cross-sectional median volatility each period.

reflected in the price of a high volatility stock.

3.8 Comparison Versus Dictionary Methods and RavenPack

Our last set of analyses compare SSESTM to alternative sentiment scoring methods in terms of return prediction accuracy.

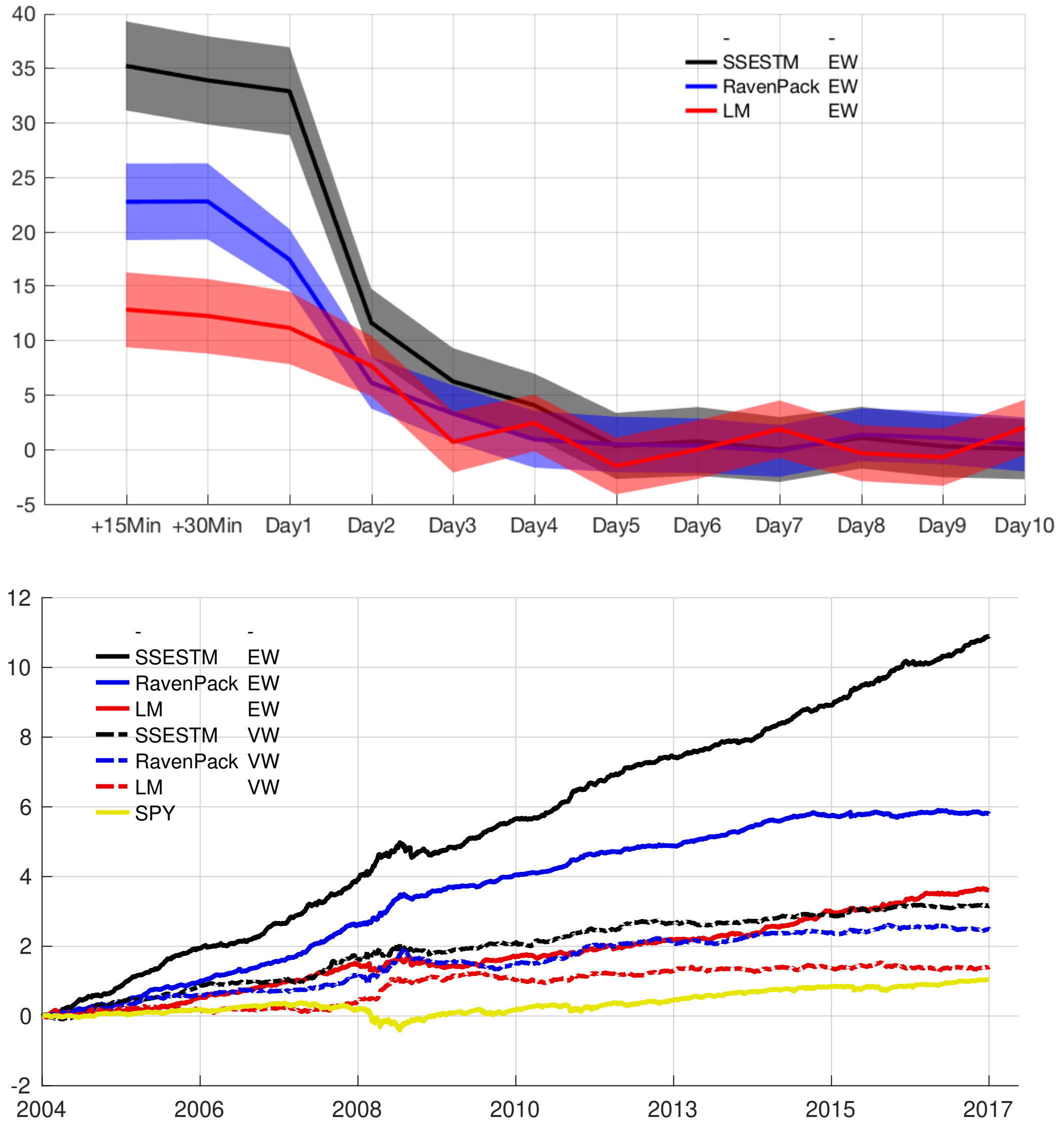
The first alternative for comparison is dictionary-based sentiment scores. We construct LM sentiment scores of a document following by aggregating counts of words listed in their positive sentiment dictionary (weighted by tf-idf, as recommended by Loughran and McDonald, 2011) and subtracting off weighted counts of words coinciding with their negative dictionary. Finally, we follow the same procedure in our analysis above to average scores from multiple articles for the same firm in the same day. This produces a stock-day signal, $\hat{p}_i^{LM}$, which may be used to construct trading strategies in the same manner that we use of SSESTM-based signal, $\hat{p}_i^{SSESTM}$, in preceding analyses.

The second alternative for comparison are news sentiment scores from RavenPack News Analytics 4 (RPNA4). As stated on its website,[17]

> RavenPack is the leading big data analytics provider for financial services. Financial professionals rely on RavenPack for its speed and accuracy in analyzing large amounts of unstructured content. The company's products allow clients to enhance returns, reduce risk and increase efficiency by systematically incorporating the effects of public information in their models or workflows.

[17] https://www.ravenpack.com/about/.

Figure 12: SSESTM Versus LM and RavenPack



Note: For top panel notes, see Figure 8. In addition to SSESTM, the top panel reports trading strategy performance for sentiment measures based on RavenPack and LM. The bottom panel compares the daily cumulative returns of long-short portfolios constructed from SSESTM, RavenPack, and LM sentiment scores, separated into equal-weighted (EW, solid lines) and value-weighted (VW, dashed lines) portfolios, respectively. The yellow solid line is the S&P 500 return (SPY).

Table 4: SSESTM Versus LM and RavenPack

	Sharpe		Average	FF6+SSESTM			FF6+LM			FF6+RP		
EW/VW	Ratio	Turnover	Return	α	$t(\alpha)$	R^2	α	$t(\alpha)$	R^2	α	$t(\alpha)$	R^2
						SSESTM						
EW	4.29	94.7%	33				29	14.96	7.8%	29	14.91	4.7%
VW	1.33	91.6%	10				9	4.92	10.2%	9	4.85	10.7%
						RavenPack						
EW	3.24	95.3%	18	15	10.87	3.0%	16	11.73	3.3%			
VW	1.14	94.8%	8	7	4.22	4.3%	8	4.45	4.1%			
						LM						
EW	1.71	94.5%	12	5	3.43	7.7%				9	5.38	4.9%
VW	0.73	93.9%	5	3	2.12	2.9%				4	2.67	3.2%

Note: The table repeats the analysis of Table 2 for the equal-weighted long-short (L-S) portfolios plotted in Figure 7, as well as their long (L) and short (S) legs. Sharpe ratios are annualized, while returns and alphas are reported in basis points per days.

> RavenPack's clients include the most successful hedge funds, banks, and asset managers in the world.

We use data from the RPNA4 DJ Edition Equities, which constructs news sentiment scores from company-level news content sourced from the same Dow Jones sources that we use to build SSESTM (*Dow Jones Newswires*, *Wall Street Journal*, *Barron's* and *MarketWatch*), thus the collection of news articles that we have access to is presumably identical to that underlying RavenPack. However, the observation count that we see in RavenPack is somewhat larger than the number of observations we can construct from the underlying Dow Jones news. We discuss this point, along with additional details of the RavenPack data, in Appendix E. Following the same procedure used for $\hat{p}_i^{SSESTM}$ and $\hat{p}_i^{LM}$, we construct RavenPack daily stock-level sentiment scores ($\hat{p}_i^{RP}$) by averaging the all reported article sentiment scores pertaining to a given firm in a given day.[18]

We build trading strategies using each of the three sentiment scores, $\hat{p}_i^{SSESTM}$, $\hat{p}_i^{LM}$, and $\hat{p}_i^{RP}$. Our portfolio formation procedure is identical to that in previous sections, buying the 50 stocks with the most positive sentiment each day and shorting the 50 with the most negative sentiment. We consider equal-weighted and value-weighted strategies.

The top panel of Figure 4 assesses the extent and timing of prices respond to each sentiment measure. It reports the average daily equally weighted trading strategy return to buying stocks with positive news sentiment and selling those with negative news sentiment. The first and most important conclusion from this figure is that SSESTM is significantly more effective than alternatives in identifying price-relevant content of news articles. Beginning fifteen minutes of news arrival, the one-day long-short return based on SSESTM is on average 33 basis points, versus 18 basis points for RavenPack and 12 for LM. The plot also shows differences in the horizons over which prices respond to each measure. The RavenPack and LM signals are fully incorporated into prices within two days

[18] We use RavenPack's flagship measure, the composite sentiment score, or CSS.

(the effect of RavenPack is borderline insignificant at three days). The SSESTM signal, on the other hand, requires four days to be fully incorporated in prices. This suggests that SSESTM is able to identify more complex information content in news articles that investors cannot fully act on within the first day or two of trading.

The bottom panel of Figure 4 focuses on the one-day trading strategy and separately analyzes equal and value weight strategies. It reports out-of-sample cumulative daily returns to compare average strategy slopes and drawdowns. This figure illustrates an interesting differentiating feature of SSESTM versus RavenPack. Following 2008, and especially in mid 2014, the slope of the RavenPack strategy noticeably flattens. While we do not have data on their subscriber base, anecdotes from the asset management industry suggest that subscriptions to RavenPack by financial institutions grew rapidly over this time period. In contrast, the slope of SSESTM is generally stable during our test sample.

Another important overall conclusion from our comparative analysis is that all sentiment strategies show significant positive out-of-sample performance. Table 4 reports a variety of additional statistics for each sentiment trading strategy including annualized Sharpe ratios of the daily strategies shown in Figure 4, as well as their daily turnover. The SSESTM strategy dominates not only in terms of average returns, but also in terms of Sharpe ratio, and with slightly less turnover than the alternatives. In equal-weighted terms, SSESTM earns an annualized Sharpe ratio of 4.3, versus 3.2 and 1.7 for RavenPack and LM, respectively. The outperformance of SSESTM is also evident when comparing value-weighted Sharpe ratios. In this case, SSESTM achieves a Sharpe ratio of 1.3 versus 1.1 for RavenPack and 0.7 for LM.

The gain in performance of SSESTM versus LM derives from their differences in scoring strategies. Our assessment of sentiment-charged words and the corresponding sentiment weights are entirely data-driven, whereas the LM dictionary is subjective and their choices of words and sentiment scores is ad hoc. SSESTM's outperformance versus RavenPack is further bolstered by the advantage that SSESTM is an interpretable white-box with theoretical guarantees, while RavenPack scores are entirely black-box.

To more carefully assess the differences in performance across methods, Table 4 reports a series of portfolio spanning tests. For each sentiment-based trading strategy, we regress its returns on the returns of each of the competing strategies, while also controlling for daily returns to the five Fama-French factors plus the UMD momentum factor (denoted FF6 in the table). We evaluate both the R^2 the regression intercept (α). If a trading strategy has a significant α after controlling for an alternative, it indicates that the underlying sentiment measure isolates predictive information that is not fully subsumed by the alternative. Likewise, the R^2 measures the extent to which trading strategies duplicate each other.

An interesting result of the spanning tests is the overall low correlation among strategies as well as with the Fama-French factors. The highest R^2 we find is 10.7% for SSESTM regressed on FF6 and the RavenPack strategy. The SSESTM α's are in each case almost as large as its raw return—at most 15% of the SSESTM strategy performance is explained by the controls (i.e., an equal-weighted α of 29 basis points versus the raw average return of 33 basis points). We also see significant positive alphas

Table 5: Performance of the Long-Short Portfolios Net Transaction Costs

		Gross		Net	
γ	Turnover	Return	Sharpe Ratio	Return	Sharpe Ratio
0.9	92.90%	33	4.29	14	1.81
0.8	90.60%	33	4.24	15	1.85
0.7	88.00%	33	4.24	15	1.93
0.6	85.30%	32	4.17	15	1.94
0.5	82.50%	32	4.08	15	1.93
0.4	79.50%	31	3.91	15	1.84
0.3	76.20%	30	3.73	14	1.78
0.2	72.30%	26	3.27	12	1.45
0.1	67.60%	22	2.58	8	0.92

Note: The table reports the performance of equally-weighted long-short portfolios constructed using exponentially smoothed sentiment scores via SSESTM. The exponential smoothing parameter is γ. The average returns and the Sharpe ratios are annualized. The portfolios' average daily turnover is calculated as Turnover $= \frac{1}{2T}\sum_{t=1}^{T}\left(\sum_i |w_{i,t+1} - w_{i,t}(1 + r_{i,t+1})|\right)$. Average returns are reported in basis points per day.

for the alternative strategies after controlling for SSESTM, indicating not only that they achieve significant positive returns, but also that a component of those excess returns are uncorrelated with SSESTM and FF6. In short, SSESTM, RavenPack, and LM capture different varieties of information content in news articles, which suggests that are potential mean-variance gains from combining the three strategies. Indeed, and portfolio that places a one-third weight on each of the equally-weighted sentiment strategies earns an annualized out-of-sample Sharpe ratio of 4.9, significantly exceeding the 4.3 Sharpe ratio of SSESTM on its own.

3.9 Transaction Costs

Our trading strategy performance analysis thus far ignores transaction costs because the portfolios are used primarily to give economic context and a sense of economic magnitude to the strength of the predictive content of each sentiment measure. The profitability of the trading strategy net of costs is neither here nor there for assessing sentiment predictability. Furthermore, the comparative analysis of SSESTM, LM, and RavenPack is apples-to-apples in the sense that all three strategies face the same trading cost environment.

That said, ascertaining the usefulness of news article sentiment for practical portfolio choice is an interesting separate question; however, this is difficult to assess from preceding tables due to the large turnover incurred by sentiment strategies. In this section, to better understand the relevance of SSESTM's predictability gains for practical asset management, we investigate the performance of sentiment-based trading strategies while taking into account trading costs.

To approximate the net performance of a strategy, we assume that each portfolio incurs a daily transaction cost of 2×turnover×10bps. That is, each unit of turnover incurs a total cost of 20bps, paid as 10bps upon entry and another 10bps upon exit of a position. The choice of 10bps approximates the average trading cost incurred by large asset managers, as demonstrated by Frazzini et al. (2018).

Our proposed adjustment is applicable to the strategies studied above, but does not recognize the fact that managers may take action to reduce trading costs. In order to address this, we consider variations of our baseline strategy that are designed to cut back portfolio turnover. To do so, we smooth sentiment scores by calculating moving averages of the history of scores at the stock-level. We opt for an exponentially-weighted moving average (EWMA) scheme, which can be represented with the recursive formula

$$\tilde{p}_{i,t} = \gamma \hat{p}_{i,t} + (1 - \gamma)\tilde{p}_{i,t-1}.$$

EWMA uses moving average weights that exponentially decay with time lag, thus placing the heaviest emphasis on the most recent observations, but still using the entire history of data. Smoothing the signal in this way is beneficial for portfolio construction if the reduction in turnover (and hence transaction costs) more than offsets any deterioration in predictive performance of the signal.[19] We vary the degree of smoothness from none at all ($\gamma = 1$) to very heavily smoothed ($\gamma = 0.05$).

Table 5 reports the performance of SSESTM portfolios based on smoothed sentiment scores. Moving down the rows we see that more smoothing naturally translates into lower turnover. At the same time, the gross Sharpe ratio of the trading strategy decreases, indicating a loss in predictive information due to smoothing. The net Sharpe ratio, however, peaks at 1.94 when $\gamma = 0.6$. Evidently, with a moderate degree of smoothing, the gain from reducing transaction costs outstrips the loss in predictive power. Thus, trading on news sentiment remains profitable after accounting for transactions costs, and some these transaction costs can be counteracted by smoothing the sentiment signal with EWMA.

4 Conclusion

We propose and analyze a new text-mining methodology, SSESTM, for extraction of sentiment information from text documents through supervised learning. In contrast to common sentiment scoring approach in the finance literature, such as dictionary methods and commercial vendor platforms like RavenPack, our framework delivers customized sentiment scores for individual research applications. This includes isolating a list of application-specific sentiment terms, assigning sentiment weights to these words via topic modeling, and finally aggregating terms into document-level sentiment scores. Our methodology has the advantage of being entirely "white box" and thus clearly interpretable, and we derive theoretical guarantees on the statistical performance of SSESTM under minimal assumptions. It is easy to use, requiring only basic statistical tools such as penalized regression, and its low computational cost makes it ideally suited for analyzing big data.

To demonstrate the usefulness of our method, we analyze the informational content of *Dow Jones Newswires* in the practical problem of portfolio construction. In this setting, our model selects lists of positive and negative words that are unmistakably coherent gauges of sentiment. The resulting news sentiment scores are powerful predictors for price responses to new information. To quantify the

[19] Note, it is not necessarily the case that smoothing $\hat{p}_{i,t}$ weakens its predictive performance. If the true sentiment is persistent and the estimated sentiment is sufficiently noisy, then time series smoothing can potentially improve the predictive signal.

economic magnitude of their predictive content, we construct simple trading strategies that handily outperform sentiment metrics from a commercial vendor widely-used in the asset management industry. We also demonstrate how our approach can be used to investigate the process of price formation in response to news.

While our empirical application targets information in business news articles for the purpose of portfolio choice, the method is entirely general. It may be adapted to any setting in which a final explanatory or forecasting objective supervise the extraction of conditioning information from a text data set.

References

Antweiler, Werner, and Murray Z Frank, 2005, Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, *The Journal of Finance* 59, 1259–1294.

Blei, David M, Andrew Y Ng, and Michael I Jordan, 2003, Latent Dirichlet Allocation, *Journal of Machine Learning Research* 3, 993–1022.

Cowles, Alfred, 1933, Can stock market forecasters forecast?, *Econometrica: Journal of the Econometric Society* 309–324.

Fama, Eugene F, 1970, Efficient capital markets: A review of theory and empirical work, *The Journal of Finance* 25, 383–417.

Fan, Jianqing, and Jinchi Lv, 2008, Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, 849–911.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, 2017, Taming the factor zoo, Technical report, University of Chicago.

Frazzini, Andrea, Ronen Israel, and Tobias J Moskowitz, 2018, Trading costs, *Working Paper* .

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber, 2017, Dissecting characteristics nonparametrically, Technical report, University of Wisconsin-Madison.

Genovese, Christopher R, Jiashun Jin, Larry Wasserman, and Zhigang Yao, 2012, A comparison of the lasso and marginal regression, *Journal of Machine Learning Research* 13, 2107–2143.

Gentzkow, Matthew, Bryan T Kelly, and Matt Taddy, forthcoming, Text as data, *Journal of Economic Literature* .

Gentzkow, Matthew, Jesse M Shapiro, and Matt Taddy, 2019, Measuring group differences in high-dimensional choices: Method and application to congressional speech, *Econometrica* .

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2018, Empirical asset pricing via machine learning, Technical report, University of Chicago.

Hofmann, Thomas, 1999, Probabilistic latent semantic analysis, in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 289–296, Morgan Kaufmann Publishers Inc.

Huang, Allen H, Amy Y Zang, and Rong Zheng, 2014, Evidence on the Information Content of Text in Analyst Reports, *The Accounting Review* 89, 2151–2180.

James, William, and Charles Stein, 1961, Estimation with quadratic loss, in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, 361–379.

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

Ji, Pengsheng, and Jiashun Jin, 2012, UPS delivers optimal phase diagram in high-dimensional variable selection, *The Annals of Statistics* 40, 73–103.

Ke, Zheng Tracy, and Minzhe Wang, 2017, A new svd approach to optimal topic estimation, Technical report, Harvard University.

Kelly, Bryan, Seth Pruitt, and Yinan Su, 2017, Some characteristics are risk exposures, and the rest are irrelevant, Technical report, University of Chicago.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2017, Shrinking the cross section, Technical report, University of Michigan.

Li, Feng, 2010, The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach, *Journal of Accounting Research* 48, 1049–1102.

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

Loughran, Tim, and Bill Mcdonald, 2016, Textual Analysis in Accounting and Finance: A Survey, *Journal of Accounting Research* 54, 1187–1230.

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

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, 2013, Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems*, 3111–3119.

Reed, Adam V, 2007, Costly short-selling and stock price adjustment to earnings announcements, *Working Paper* .

Santosh, Shrihari, 2016, The speed of price discovery: Trade time vs clock time, *Working Paper* .

Shorack, Galen R, and Jon A Wellner, 2009, *Empirical processes with applications to statistics*, volume 59 (Siam).

Tetlock, Paul C, 2007, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, *The Journal of Finance* 62, 1139–1168.

Tetlock, Paul C, Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More Than Words: Quantifying Language to Measure Firms' Fundamentals, *Journal of Finance* 63, 1437–1467.

Veldkamp, Laura L, 2006, Information markets and the comovement of asset prices, *The Review of Economic Studies* 73, 823–845.

Wilson, Robert, 1975, Informational economies of scale, *The Bell Journal of Economics* 184–195.

Appendix

A Algorithms

Algorithm 1.

S1. For each word $1 \leq j \leq m$, let

$$f_j = \frac{\#\text{ articles including word } j \text{ AND having } \mathrm{sgn}(y) = 1}{\#\text{ articles including word } j}$$

S2. For a proper threshold $\alpha_+ > 0$, $\alpha_- > 0$, and $\kappa > 0$ to be determined, construct

$$\widehat{S} = \big\{j : f_j \geq 1/2 + \alpha_+\big\} \cup \big\{j : f_j \leq 1/2 - \alpha_-\big\} \cap \{j : k_j \geq \kappa\},$$

where k_j is the total count of articles in which the jth word appears.

Algorithm 2.

S1. Sort the returns $\{y_i\}_{i=1}^n$ in the ascending order. For each $1 \leq i \leq n$, let

$$\widehat{p}_i = \frac{\text{rank of } y_i \text{ in all returns}}{n}. \tag{A.1}$$

S2. For $1 \leq i \leq n$, let $\widehat{s}_i$ be the total counts of words from $\widehat{S}$ in article i, and let $\widehat{d}_i = \widehat{s}_i^{-1} d_{i,[\widehat{S}]}$. Write $\widehat{D} = [\widehat{d}_1, \widehat{d}_2, \dots, \widehat{d}_n]$. Construct

$$\widehat{O} = \widehat{D}\widehat{W}'(\widehat{W}\widehat{W}')^{-1}, \qquad \text{where} \quad \widehat{W} = \begin{bmatrix} \widehat{p}_1 & \widehat{p}_2 & \cdots & \widehat{p}_n \\ 1-\widehat{p}_1 & 1-\widehat{p}_2 & \cdots & 1-\widehat{p}_n \end{bmatrix}. \tag{A.2}$$

Set negative entries of $\widehat{O}$ to zero and re-normalize each column to have a unit ℓ^1-norm. We use the same notation $\widehat{O}$ for the resulting matrix.

Algorithm 3.

S1. Let $\widehat{s}$ be the total count of words from $\widehat{S}$ in the new article. Obtain $\widehat{p}$ by

$$\widehat{p} = \arg\max_{p \in [0,1]} \left\{ \widehat{s}^{-1} \sum_{j=1}^{\widehat{s}} d_j \log\Big(p\widehat{O}_{+,j} + (1-p)\widehat{O}_{-,j}\Big) + \lambda \log\big(p(1-p)\big) \right\}, \tag{A.3}$$

where d_j, $\widehat{O}_{+,j}$, and $\widehat{O}_{-,j}$ are the jth entries of the corresponding vectors, and $\lambda > 0$ is a tuning parameter.

B Monte Carlo Simulations

In this section, we provide Monte Carlo evidence to illustrate the finite sample performance of the estimators we propose in the algorithms above.

We assume the data generating process of the positive, negative, and neutral words in each article follow:

$$d_{i,[S]} \sim \text{Multinomial}\Big(s_i,\ p_i O_+ + (1-p_i)O_-\Big), \quad d_{i,[N]} \sim \text{Multinomial}\Big(n_i,\ O_0\Big), \tag{B.4}$$

where $p_i \sim \text{Unif}(0,1)$, $s_i \sim \text{Unif}(0, 2\bar{s})$, $n_i \sim \text{Unif}(0, 2\bar{n})$, and for $j = 1, 2, \ldots, S$,

$$O_+(j) = \frac{2}{|S|}\left(1 - \frac{j}{|S|}\right)^2 + \frac{2}{3|S|} \times 1_{\left\{j < \frac{|S|}{2}\right\}}, \quad O_-(j) = \frac{2}{|S|}\left(\frac{j}{|S|}\right)^2 + \frac{2}{3|S|} \times 1_{\left\{j \geq \frac{|S|}{2}\right\}},$$

and $O_0(j) \sim \frac{1}{m-|S|}\text{Unif}(0,2)$, for $j = |S|+1, \ldots, m$. As a result, the first $|S|/2$ words are positive, the next $|S|/2$ words are negative, and the remaining ones are neutral with frequencies randomly drawn from a uniform distribution.

Next, the sign of returns follows a logistic regression model: $\mathbb{P}(y_i > 0) = p_i$, and its magnitude $|y_i|$ follows a Student t-distribution with the degree of freedom parameter set at 4. The standard deviation of the t-distribution does not affect our simulations, since only the ranks of returns matter.

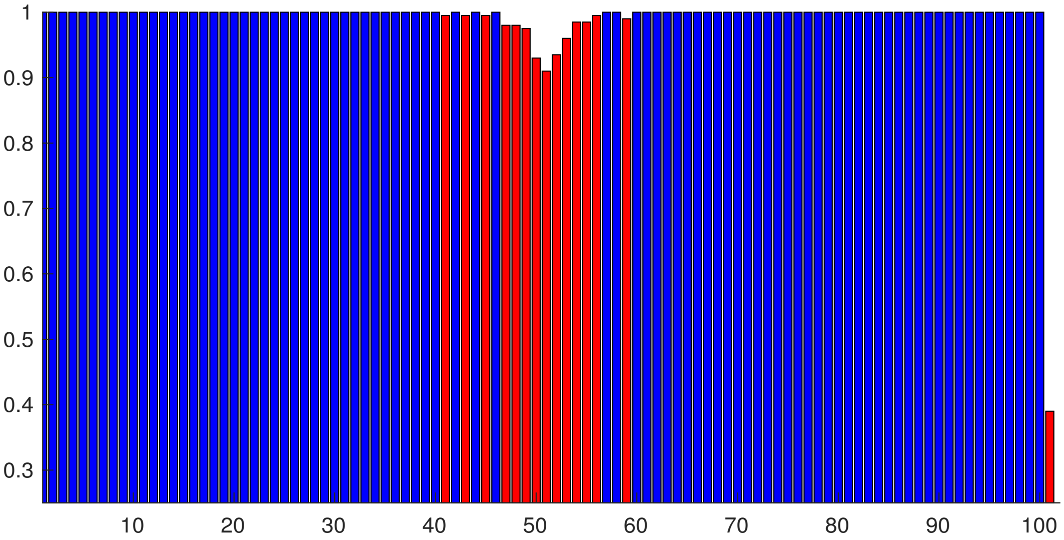
We fix the number of Monte Carlo repetitions $M_c = 200$ and the number of articles in the testing sample is $1,000$. In the benchmark case, we select $|S| = 100$, $m = 500$, $n = 10,000$, $\bar{s} = 10$, and $\bar{n} = 100$.

We first conduct an evaluation of the screening step. Instead of tuning those threshold parameters, we select a fixed amount of words, say, $|S|$, which achieve larger values in terms of $|f_j - 0.5|1_{\{k_j > \kappa\}}$, where κ is set at the 10% quantiles of all k_js. We report in Figure A.1 the frequencies of each word selected in the screening step across all Monte Carlo repetitions. There is less than 0.4% probability of selecting any word outside the set S. Not surprisingly, the words in S that are occasionally missed are those with corresponding entires of T around 0. Such words are closer to those neural words in the set N.

Next, Figure A.2 illustrates the accuracy of the estimation step, taking into account the potential errors in the screening step. The true values of T and F are shown in black. The scaling constant $\rho \approx 0.5$ in our current setting. As shown from this plot, the estimators $\widehat{F}$ and $\widehat{T}$ are fairly close to their targets F and ρT across all words, as predicted by our theory. The largest finite sample errors in $\widehat{F}$ occur to those words in F that are occasionally missed from the screening step.

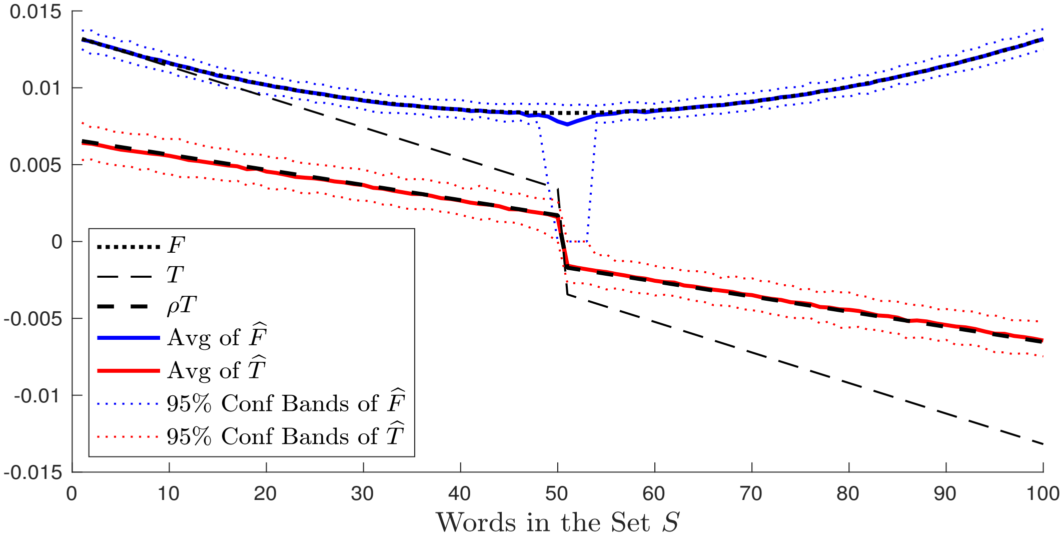
Finally, we examine the accuracy of the scoring step, with errors accumulated from the previous steps. Data of the testing sample are never used in the previous two steps. Table A.1 reports the Spearman's rank correlation coefficients between the predicted $\widehat{p}$ and the true p for 1,000 articles in the testing sample in a variety of cases. We report the rank correlation because what matters is the rank of all articles instead of their actual scores, which are difficult to consistently estimate, because of the biases in the previous steps. Also, the penalization term ($\lambda = 0.5$) in our likelihood biases the

Figure A.1: Screening Results in Simulations



Note: This figure reports the frequencies of each word in the set S selected in the screening step across all Monte Carlo repetitions. The red bars correspond to those words with frequencies less than 100%. The red bar on the right reports the aggregate frequency of a selected word outside the set S.

Figure A.2: Estimation Results in Simulations



Note: This figure compares the averages of $\widehat{F}$ (blue, solid) and $\widehat{T}$ (red, solid) across Monte Carlo repetitions with F (black, dotted), T (thin, black, dashed), and ρT (thick, black, dashed), respectively, using the benchmark parameters. The blue and red dotted lines plot the 2.5% and 97.5% quantiles of the Monte Carlo estimates.

estimated scores towards 0.5, although it has no impact on their ranks. In the benchmark setting, the average correlation across all Monte Carlo repetitions is 0.85 with a standard deviation 0.0014. If we decrease $\bar{s}$ from 10 to 5, the quality of the estimates becomes worse due to less observations from words in S. Similarly, when decrease n to 5,000, the estimates become less accurate, since the sample size is smaller. If the size of the dictionary, m, or the size of the dictionary of the sentiment words, $|S|$, drop by half, the estimates improve, despite that the improvement is marginal. Overall, these observations match what the statistical theory predicts.

Table A.1: Spearman's Correlation Estimates

	benchmark	$\bar{s}\downarrow$	$n\downarrow$	$m\downarrow$	$\|S\|\downarrow$
Avg S-Corr	0.850	0.776	0.834	0.857	0.852
Std Dev	0.0014	0.0043	0.0024	0.0025	0.0009

Note: In this table, we report the mean and standard deviation of Spearman's correlation estimates across Monte Carlo repetitions for a variety of cases. The parameters in the benchmark case are set as: $|S| = 100$, $m = 500$, $n = 10,000$, and $\bar{s} = 10$. In each of the remaining columns, the corresponding parameter is decreased by half, whereas the rest three parameters are fixed the same as the benchmark case.

C Statistical Theory

We quantify the statistical accuracy of our method in an asymptotic framework, where the number of training articles, n, and the dictionary size, m, both go to infinity. Our framework allows the average length of training articles to be finite or go to infinity, so the theory applies to both "short" and "long" articles in the training sample. Without loss of generality, we consider a slightly different screening procedure:

$$\widehat{S} = \{j : f_j \geq 1/2 + \alpha_+\} \cup \{j : f_j \leq 1/2 - \alpha_-\}. \tag{C.5}$$

where

$$f_j = \frac{\text{count of word } j \text{ in articles with } \operatorname{sgn}(y) = +1}{\text{count of word } j \text{ in all articles}}.$$

It has rather similar theoretical properties as the screening procedure in Section 2, but the conditions and conclusions are more elegant and transparent, so we choose to present theory using this approach. The approach in the main text has a better empirical performance because of more tuning parameters.

C.1 Regularity Conditions

Let $s_{\max}$, $s_{\min}$, and $\bar{s}$ be the maximum, minimum, and average of $\{s_i\}_{i=1}^n$, respectively. In our model, for sentiment-neutral words, $d_{i,[N]}$ follows a multinomial distribution. Define $\Omega_i = \mathbb{E}d_{i,[N]}$.[20] For each $j \in N$, let $\Omega_{\min,j}$, $\Omega_{\max,j}$, and $\bar{\Omega}_{\cdot,j}$ be the maximum, minimum, and average of $\{\Omega_{i,j}\}_{i=1}^n$, respectively.

[20] If we write $d_{i,[N]} \sim \text{Multinomial}(n_i, q_i)$, where n_i is the total count of words from N in document i and $q_i \in \mathbb{R}_+^{|N|}$ is a distribution on the space of N, then $\Omega_i = n_i q_i$.

We assume

$$\frac{s_{\max}}{\bar{s}} \leq C, \qquad \max_{j\in N} \frac{\Omega_{\max,j}}{\bar{\Omega}_{\cdot,j}} \leq C, \qquad \min_{j\in S} \frac{n\bar{s}(O_{+,j}+O_{-,j})}{\log(m)} \to \infty, \qquad \min_{j\in N} \frac{n\bar{\Omega}_{\cdot,j}}{\log(m)} \to \infty. \tag{C.6}$$

The last two inequalities in (C.6) require the expected count of any word in all of n training articles to be much larger than $\log(m)$. Since n is large in real data, this condition is mild. For a constant $c_0 \in (0,1)$, we assume

$$\min_{j\in S} \frac{\sum_{i=1}^n s_i \left[p_i O_{+,j} + (1-p_i)O_{-,j}\right]}{\sum_{i=1}^n s_i(O_{+,j}+O_{-,j})} \geq c_0. \tag{C.7}$$

This condition says that the expected count of a word $j \in S$ in all training articles cannot be much smaller than $n\bar{s}F_j$, where F_j is the vector of frequency defined in (8). It is for technical convenience. We also assume

$$\frac{1}{n}\sum_{i=1}^n p_i = \frac{1}{2}, \qquad \frac{\sum_{i=1}^n s_i \, \mathbb{E}[\mathrm{sgn}(y_i)]}{\sum_{i=1}^n s_i} = 0, \tag{C.8}$$

This condition essentially requires that we have approximately equal number of articles with positive and negative tone. Note that we can always keep the same number of articles associated with positive and negative returns in the training stage, so this condition is mild. We also assume

$$\frac{\sum_{i=1}^n \Omega_{i,j} \, \mathbb{E}[\mathrm{sgn}(y_i)]}{\sum_{i=1}^n \Omega_{i,j}} = 0, \qquad \text{for all } j \in N. \tag{C.9}$$

This condition ensures that the count of any sentiment-neutral word has no correlation with the sign of the stock returns (so they are indeed "sentiment-neutral"). All equalities in (C.8)-(C.9) do not need to hold exactly. We impose exact equalities so that the conclusions are more elegant.

C.2 Accuracy of the Estimators in Algorithms 1 and 2

First, we consider the screening step. We define a quantity to capture the *sensitivity of stock returns to article sentiment*:

$$\theta \equiv \frac{\sum_{i=1}^n s_i\left(p_i - \frac{1}{2}\right)\left[g(p_i) - \frac{1}{2}\right]}{\sum_{i=1}^n s_i}, \tag{C.10}$$

where $g(\cdot)$ is the monotone increasing function defined in (2). When $g(\frac{1}{2}) = \frac{1}{2}$, this quantity is lower bounded by $[\min_{x\in[0,1]} g'(x)][\frac{1}{n\bar{s}}\sum_{i=1}^n s_i(p_i - \frac{1}{2})^2]$. Roughly speaking, θ measures the steepness of g and the extremeness of training articles' polarities.

Theorem C.1. *Consider the model (1)-(4), where (C.6)-(C.9) hold. As $n, m \to \infty$, with probability $1 - o(1)$,*

$$|f_j - 1/2| \begin{cases} \geq 2\theta \frac{|O_{+,j}-O_{-,j}|}{O_{+,j}+O_{-,j}} + \frac{C\sqrt{\log(m)}}{\sqrt{n\min\{1, \bar{s}(O_{+,j}+O_{-,j})\}}}, & \text{for } j \in S, \\ \leq \frac{C\sqrt{\log(m)}}{\sqrt{n\min\{1, \bar{\Omega}_{\cdot,j}\}}}, & \text{for } j \in N. \end{cases}$$

The set of retrained words, $\widehat{S}$, is obtained by thresholding $|f_j - 1/2|$ at $\alpha_{\pm}$. Theorem C.1 suggests that $|f_j - 1/2|$ is large for sentiment-sensitive words and small for sentiment-neutral words, justifying that the screening step is meaningful. We say that the screening step has the *sure-screening property* (Fan and Lv, 2008) if $\mathbb{P}(\widehat{S} = S) = 1 - o(1)$.

Theorem C.2 (Sure Screening)**.** *Consider the model* (1)-(4)*, where* (C.6)-(C.9) *hold. We assume*

$$n\theta^2 \min_{j \in S} \frac{(O_{+,j} - O_{-,j})^2}{(O_{+,j} + O_{-,j})^2} \geq \frac{\log^2(m)}{\min\{1,\ \bar{s}\min_{j\in S}(O_{+,j} + O_{-,j}),\ \min_{j\in N} \bar{\Omega}_{\cdot,j}\}}. \tag{C.11}$$

In the screening step (C.5)*, we set* $\alpha_\pm = \frac{\sqrt{\log(m)\log(\log(m))}}{\sqrt{n\min\{1, \bar{s}\min_{j\in S}(O_{+,j}+O_{-,j}), \min_{j\in N}\bar{\Omega}_{\cdot,j}\}}}$*. Then, as* $n, m \to \infty$*,* $\mathbb{P}(\widehat{S} = S) = 1 - o(1)$*.*

The desired number of training articles for sure screening is determined by three factors. First, θ. The sensitivity of stock returns to article sentiment, defined in (C.10). Second, $\min_{j\in S}\frac{|O_{+,j}-O_{-,j}|}{O_{+,j}+O_{-,j}}$. It represents the word's frequency-adjusted sentiment. Third, $\min\{1, \bar{s}\min_{j\in S}(O_{+,j}+O_{-,j}), \min_{j\in N}\bar{\Omega}_{\cdot,j}\}$. Note that the last two terms in the minimum are related to the per-article count of individual words. For "long articles" where the per-article count of each word is bounded below by a constant, this factor equals 1. For "short articles", the per-article count of a word may tend to zero, so we need to have more training articles.

Next, we consider the estimation step of Algorithm 2. We quantify the estimation errors on F and T. The results can be directly translated to estimation errors on O_+ and O_-.

Theorem C.3 (Estimation Error of Sentiment Vectors)**.** *Consider the model* (1)-(4)*, where* (C.6)-(C.9) *and* (C.11) *hold. As* $n, m \to \infty$*, with probability* $1 - o(1)$*,*

$$\|\widehat{F} - F\|_1 \leq C\sqrt{\frac{|S|\log(m)}{n\bar{s}}}, \qquad \|\widehat{T} - \rho T\|_1 \leq C\sqrt{\frac{|S|\log(m)}{n\bar{s}}}.$$

We now compare the rate with the theoretical results of topic estimation in unsupervised settings. It was shown in Ke and Wang (2017) that, given n articles, written on a size-$|S|$ dictionary, with an average length of $\bar{s}$, the minimax convergence rate of the ℓ^1-norm between true and estimated topic vectors is

$$\sqrt{\frac{|S|}{n\bar{s}}}, \qquad \text{up to a logarithmic factor.}$$

Our model imposes a 2-topic topic model on sentiment-sensitive words, so the intrinsic discionary size is $|S|$. Therefore, our method has achieved the best possible error rate of unsupervised methods. However, for unsupervised methods to achieve this rate, they typically require the average document length to be much larger than the dictionary size (Ke and Wang, 2017). Translated to our setting, it means the total count of sentiment-sensitive words in one article needs to be much larger than the number of sentiment-sensitive words. This is not satisfied in our empirical study, where the identified sentiment dictionary has $100 \sim 200$ words, yet their total count in one article is typically below 20. In this case, our supervised approach has a much smaller error rate than the unsupervised methods.

However, the supervised approach comes with a price: Our method is estimating $(F, \rho T)$, instead of (F, T). Note that our assumption (2) ensures $\rho > 0$. It means, regardless of the errors of estimating

p_i by $\widehat{p}_i$, our method always *preserves the order of the tone of words*. This property is very important, as it guarantees that in the scoring step we always correctly identify whether a new article has positive or negative sentiment, regardless of the errors in $\widehat{p}_i$.

When $\widehat{p}_i = p_i$, the factor $\rho = 1$. So, our method precisely estimates T. When $\widehat{p}_i \neq p_i$, this factor is smaller than 1, so our method "discounts" the vector of tone. Once the exact distribution of y_i given p_i is specified, this factor can be computed explicitly.

C.3 Accuracy of the Estimator in Algorithm 3

Given a new article with sentiment p, define the *rescaled sentiment* as

$$p^* = \frac{1}{2} + \rho^{-1}\Big(p - \frac{1}{2}\Big).^{21} \tag{C.12}$$

It maps $p \in [0,1]$ to $p^* \in [\frac{1-\rho^{-1}}{2}, \frac{1+\rho^{-1}}{2}]$, while preserving the order of $(p - \frac{1}{2})$. Our scoring step gives a consistent estimator of p^*.

Theorem C.4 (Scoring Error on New Article)**.** *Consider the model (1)-(4), where (C.6)-(C.9) hold. Define $O^{(\rho)} = [O_+^{(\rho)}, O_-^{(\rho)}]$, with $O_\pm^{(\rho)} = F \pm \rho T$. Suppose (C.11) is satisfied with O replaced by $O^{(\rho)}$. Let $d \in \mathbb{R}_+^m$ be the word count vector of a new article with sentiment p. For a constant $c_1 \in (0, \frac{1}{2})$, we assume that $pO_{+,j} + (1-p)O_{-,j} \geq c_1(O_{+,j} + O_{-,j})$, for all $j \in S$, and that $c_1 \leq p^* \leq 1 - c_1$, where p^* is the rescaled sentiment. Write*

$$err_n = \frac{1}{\rho\sqrt{\Theta}}\left(\frac{\sqrt{|S|\log(m)}}{\rho\sqrt{n\bar{s}\Theta}} + \frac{1}{\sqrt{s}}\right), \qquad \textit{where} \quad \Theta = \sum_{j\in S}\frac{(O_{+,j} - O_{-,j})^2}{O_{+,j} + O_{-,j}}.$$

We assume the length of the new article satisfies $s\Theta \to \infty$. Let $\widehat{p}$ be the estimator in (A.3) with a tuning parameter $\lambda > 0$. For any $\epsilon > 0$, with probability $1 - \epsilon$,

$$|\widehat{p} - p^*| \leq C\min\Big\{1, \frac{\rho^2\Theta}{\lambda}\Big\}err_n + C\min\Big\{1, \frac{\lambda}{\rho^2\Theta}\Big\}|p^* - \frac{1}{2}|.$$

Therefore, the optimal choice of tuning parameter is $\lambda = \frac{\rho^2\Theta}{|p^* - \frac{1}{2}|}err_n$, and the associated scoring error is $|\widehat{p} - p^*| \leq C\min\{err_n, |p^* - \frac{1}{2}|\}$.

The choice of λ yields a bias-variance trade-off. In the error bound for $|\widehat{p} - p^*|$, the first term $\min\{1, \frac{\rho^2\Theta}{\lambda}\}err_n$ is the "variance" term, decreasing with λ; the second term $\min\{1, \frac{\lambda}{\rho^2\Theta}\}|p^* - \frac{1}{2}|$ is the "bias" term, increasing with λ. In reality, it is a common belief that the majority of articles have a neutral tone, so the bias is negligible. At the same time, text data are very noisy, so adding the penalty can significantly reduce the variance. Our estimator shares the same spirit as the James-Stein estimator (James and Stein, 1961) by shrinking the MLE of p towards $\frac{1}{2}$. Interestingly, given that

[21] In this subsection, we condition on the returns $\{y_i\}_{i=1}^n$ and let the probability law be with respective to only the randomness of word counts in training and testing articles. By (1), the probability laws with and without conditioning are the same. Since the randomness of ρ is from $\{y_i\}_{i=1}^n$, through out this subsection, ρ is treated as non-random.

the true sentiment p is closer to $\frac{1}{2}$ than p^*, the shrinkage effect here helps reduce the scaling effect in (C.12), which means in some scenarios our estimator does a better job estimating the original p.

The error rate err_n has two terms, corresponding to the noise level in the training phase and the scoring phase, respectively. Since n is large, the latter always dominates. To guarantee $err_n \to 0$, we need that the length of the new article goes to infinity asymptotically. Nonetheless, the length of training articles can be finite.

Our estimator has a bias on estimating the original sentiment p. When the estimation quality in $\hat{p}_i$'s is good, $\rho \approx 1$ and the bias $(p^* - p)$ is small. More importantly, even with a large bias, it has no impact on practical usage, as the estimator preserves the relative rank of sentiments when applied to score multiple articles.

Theorem C.5 (Rank Correlation with True Sentiment)**.** *Under conditions of Theorem C.4, suppose we are given N new articles whose sentiments $p_1, \ldots, p_N$ are iid sampled from a continuous distribution on $\mathcal{P}(c_1) \equiv \{p \in [0,1] : pO_{+,j} + (1-p)O_{-,j} \geq c_1(O_{+,j} + O_{-,j}), \text{for all } j \in S; c_1 \leq p^* \leq 1 - c_1\}$, where $c_1 \in (0, \frac{1}{2})$ is a constant. We assume the length of each new article i satisfies $C^{-1}s \leq s_i \leq Cs$, where $s\Theta/\sqrt{\log(N)} \to \infty$. We apply the estimator (A.3) with $\lambda = \rho^2\Theta \cdot err_n$ to score all new articles. Let $SR(\hat{p}, p)$ be the Spearman's rank correlation between $\{\hat{p}\}_{i=1}^N$ and $\{p_i\}_{i=1}^N$. As $n, m, N \to \infty$,*

$$\mathbb{E}[SR(\hat{p}, p)] \to 1.$$

D Mathematical Proofs

D.1 Proofs of Theorem C.1 and Theorem C.2

Proof. First, we prove Theorem C.1. For each word $1 \leq j \leq m$, let L_j^+ and L_j^- be the total counts of word j in articles with positive and negative returns, respectively. Write for short $t_i = \text{sgn}(y_i) \in \{\pm 1\}$, for $1 \leq i \leq n$. Then, $L_j^{\pm} = \sum_{i=1}^n \frac{1 \pm t_i}{2} \cdot d_{i,j}$. It follows that

$$f_j = \frac{1}{2} + \frac{1}{2}\frac{L_j^+ - L_j^-}{L_j^+ + L_j^-} = \frac{1}{2} + \frac{\sum_{i=1}^n t_i \cdot d_{i,j}}{\sum_{i=1}^n d_{i,j}}. \tag{D.13}$$

Below, we study f_j for $j \in S$ and $j \in N$, separately.

Consider $j \in S$. As in (8), we let $F = \frac{1}{2}(O_+ + O_-)$ and $T = \frac{1}{2}(O_+ - O_-)$. We also introduce the notations $\eta_i = 2p_i - 1$ and $\eta_i(g) = 2g(p_i) - 1$. By our model, $d_i \sim \text{Multinomial}\big(s_i,\ p_iO_+ + (1-p_i)O_-\big)$, where $p_iO_+ + (1-p_i)O_- = \frac{1+\eta_i}{2}O_+ + \frac{1-\eta_i}{2}O_- = F + \eta_i T$. It follows that

$$d_{i,j} \sim \text{Binomial}\big(s_i,\ F_j + \eta_i T_j\big). \tag{D.14}$$

Let $\{b_{i,j,\ell}\}_{\ell=1}^{s_i}$ be a collection of *iid* Bernoulli variables with a success probability $(F_j + \eta_i T_j)$. Then, $d_{i,j} \overset{(d)}{=} \sum_{\ell=1}^{s_i} b_{i,j,\ell}$, where $\overset{(d)}{=}$ means two variables have the same distribution. It follows that

$$f_j \overset{(d)}{=} \frac{1}{2} + \frac{\sum_{i=1}^n \sum_{\ell=1}^{s_i} t_i \cdot b_{i,j,\ell}}{\sum_{i=1}^n \sum_{\ell=1}^{s_i} b_{i,j,\ell}}, \qquad \text{where} \quad b_{i,j,\ell} \overset{iid}{\sim} \text{Bernoulli}(F_j + \eta_i T_j). \tag{D.15}$$

The variables $\{b_{i,j,\ell}\}$ are mutually independent, with $|b_{i,j,\ell}| \leq 1$, $\mathbb{E}b_{i,j,\ell} = F_j + \eta_i T_j$ and $\mathrm{var}(b_{i,j,\ell}) \leq F_j + \eta_i T_j \leq 2F_j$. Using the Bernstein's inequality (Shorack and Wellner, 2009), we obtain that, with probability $1 - O(m^{-2})$,

$$\begin{aligned}\Big|\sum_{i=1}^n \sum_{\ell=1}^{s_i} b_{i,j,\ell} - \sum_{i=1}^n s_i(F_j + \eta_i T_j)\Big| &\leq C\sqrt{\sum_{i=1}^n 2s_i F_j \log(m)} + \log(m) \\ &\leq C\sqrt{n\bar{s}F_j \log(m)} + \log(m) \\ &\leq C\sqrt{n\bar{s}F_j \log(m)},\end{aligned}$$

where the last inequality is due to (C.6) which says $n\bar{s}F_j \gg \log(m)$. Similarly, we apply Bernstein's inequality to study $\sum_{i=1}^n \sum_{\ell=1}^{s_i} t_i \cdot q_{i,j,\ell}$. By our model (1), $\{t_i\}_{i=1}^n$ and $\{d_{i,j}\}_{1\leq i\leq n, 1\leq j\leq m}$ are mutually independent. We thereby condition on $\{t_i\}_{i=1}^n$. It follows that, with probability $1 - O(m^{-2})$,

$$\Big|\sum_{i=1}^n \sum_{\ell=1}^{s_i} t_i \cdot b_{i,j,\ell} - \sum_{i=1}^n t_i \cdot s_i(F_j + \eta_i T_j)\Big| \leq C\sqrt{n\bar{s}F_j \log(m)}.$$

We plug the above inequalities into (D.15). It gives

$$\begin{aligned}f_j &= \frac{1}{2} + \frac{\sum_{i=1}^n t_i s_i(F_j + \eta_i T_j) + O\big(\sqrt{n\bar{s}F_j \log(m)}\big)}{\sum_{i=1}^n s_i(F_j + \eta_i T_j) + O\big(\sqrt{n\bar{s}F_j \log(m)}\big)} \\ &= \frac{1}{2} + \frac{F_j \sum_{i=1}^n t_i s_i + T_j \sum_{i=1}^n t_i \eta_i s_i + O\big(\sqrt{n\bar{s}F_j \log(m)}\big)}{F_j \sum_{i=1}^n s_i + T_j \sum_{i=1}^n \eta_i s_i + O\big(\sqrt{n\bar{s}F_j \log(m)}\big)}.\end{aligned} \tag{D.16}$$

In the denominator, the sum of the first two terms can be rewritten as $\sum_{i=1}^n s_i[p_i O_{+,j} + (1-p_i)O_{-,j}]$. It is upper bounded by $2n\bar{s}F_j$, and by (C.7), it is also lower bounded by $2c_0 n\bar{s}F_j$. Furthermore, since $n\bar{s}F_j \gg \log(m)$, the last term is negligible compared to the first two terms. Hence, the denominator in (D.16) is between $c_0 n\bar{s}F_j$ and $4n\bar{s}F_j$. It follows that

$$|f_j - 1/2| \geq \frac{|T_j \sum_{i=1}^n t_i \eta_i s_i|}{4n\bar{s}F_j} - \frac{|F_j \sum_{i=1}^n t_i s_i|}{c_0 n\bar{s}F_j} + \frac{O\big(\sqrt{n\bar{s}F_j \log(m)}\big)}{c_0 n\bar{s}F_j} \tag{D.17}$$

We now deal with the randomness of $\{t_i\}_{i=1}^n$. They are independent variables such that $|t_i| \leq 1$ and $\mathbb{E}t_i = \eta_i(g)$. It follows that $\sum_{i=1}^n \eta_i s_i \mathbb{E}[t_i] = \sum_{i=1}^n s_i \eta_i \eta_i(g) = 4n\bar{s}\theta$ and $\sum_{i=1}^n |\eta_i s_i t_i|^2 \leq 4\sum_{i=1}^n s_i^2 \leq 4ns_{\max}\bar{s} \leq Cn\bar{s}^2$. Plugging them into the Hoeffding's inequality (Shorack and Wellner, 2009) gives: with probability $1 - O(m^{-2})$,

$$\Big|\sum_{i=1}^n \eta_i s_i t_i - 4n\bar{s}\theta\Big| \leq C\bar{s}\sqrt{n\log(m)}.$$

In particular, we know that $|\sum_{i=1}^n \eta_i s_i t_i| \geq 2n\bar{s}\theta$. Similarly, with probability $1-O(m^{-2})$, $|\sum_{i=1}^n s_i t_i - \sum_{i=1}^n s_i \mathbb{E}t_i| \leq C\bar{s}\sqrt{n\log(m)}$. Note that $\sum_{i=1}^n s_i \mathbb{E}t_i = 0$, due to the second equality in (C.8). So, we have $|\sum_{i=1}^n s_i t_i| \leq C\bar{s}\sqrt{n\log(m)}$. We plug these results into (D.17) and find out that

$$\begin{aligned} |f_j - 1/2| &\geq \frac{|T_j| 2n\bar{s}\theta}{4n\bar{s}F_j} - \frac{F_j \cdot C\bar{s}\sqrt{n\log(m)}}{c_0 n\bar{s}F_j} + \frac{O\big(\sqrt{n\bar{s}F_j\log(m)}\big)}{c_0 n\bar{s}F_j} \\ &\geq \frac{\theta|T_j|}{2F_j} + O\Big(\sqrt{\tfrac{\log(m)}{n}}\Big) + O\Big(\sqrt{\tfrac{\log(m)}{n\bar{s}F_j}}\Big). \end{aligned} \tag{D.18}$$

This gives the first claim of Theorem C.1.

Consider $j \in N$. We model that $d_{i,[N]}$ follows a multinomial distribution with $\mathbb{E}d_{i,[N]} = \Omega_i$. Equivalently, $d_{i,[N]} \sim \text{Multinomial}(k_i, q_i)$, where k_i is the count of all words from N in article i and $q_i \equiv k_i^{-1}\Omega_i$. Same as before, we view $d_{i,j}$ as the sum of k_i *iid* Bernoulli variables, each with a success probability of $q_{i,j}$. Using the Bernstein's inequality, we can prove that, with probability $1 - O(m^{-2})$, $|\sum_{i=1}^n d_{i,j} - \sum_{i=1}^n k_i q_{i,j}| \leq C\sqrt{\sum_{i=1}^n k_i q_{i,j}\log(m)} + \log(m)$. Here, $\sum_{i=1}^n k_i q_{i,j} = \sum_{i=1}^n \Omega_{i,j} = n\bar{\Omega}_{\cdot,j}$, where by (C.6), $n\bar{\Omega}_{\cdot,j} \gg \log(m)$. Therefore, we have

$$\Big|\sum_{i=1}^n d_{i,j} - \sum_{i=1}^n \Omega_{i,j}\Big| \leq C\sqrt{n\bar{\Omega}_{\cdot,j}\log(m)}.$$

Similarly, conditioning on $\{t_i\}_{i=1}^n$, with probability $1 - O(m^{-2})$,

$$\Big|\sum_{i=1}^n t_i d_{i,j} - \sum_{i=1}^n t_i\Omega_{i,j}\Big| \leq C\sqrt{n\bar{\Omega}_{\cdot,j}\log(m)}.$$

Plugging them into (D.13) gives

$$\begin{aligned} f_j &= \frac{1}{2} + \frac{\sum_{i=1}^n t_i\Omega_{i,j} + O\big([n\bar{\Omega}_{\cdot,j}\log(m)]^{\frac{1}{2}}\big)}{\sum_{i=1}^n \Omega_{i,j} + O\big([n\bar{\Omega}_{\cdot,j}\log(m)]^{\frac{1}{2}}\big)} \\ &= \frac{1}{2} + \frac{\sum_{i=1}^n t_i\Omega_{i,j} + O\big([n\bar{\Omega}_{\cdot,j}\log(m)]^{\frac{1}{2}}\big)}{n\bar{\Omega}_{\cdot,j} + O\big([n\bar{\Omega}_{\cdot,j}\log(m)]^{\frac{1}{2}}\big)}. \end{aligned} \tag{D.19}$$

We then deal with the randomness of $\{t_i\}_{i=1}^n$. By Hoeffding's inequality, with probability $1-O(m^{-2})$, $|\sum_{i=1}^n \Omega_{i,j}(t_i - \mathbb{E}t_i)| \leq C\sqrt{\sum_{i=1}^n \Omega_{i,j}^2\log(m)} \leq C\bar{\Omega}_{\cdot,j}\sqrt{n\log(m)}$, where the last inequality is from the condition $\Omega_{\max,j} \leq C\bar{\Omega}_{\cdot,j}$. Moreover, by our condition (C.8), $\sum_{i=1}^n \Omega_{i,j}\mathbb{E}t_i = 0$. The above imply

$$\Big|\sum_{i=1}^n t_i\Omega_{i,j}\Big| \leq C\bar{\Omega}_{\cdot,j}\sqrt{n\log(m)}.$$

We plug it into (D.19) and note that the denominator of (D.19) is $\gtrsim n\bar{\Omega}_{\cdot,j}$, since $n\bar{\Omega}_{\cdot,j} \gg \log(m)$. It follows that

$$|f_j - 1/2| \leq \frac{C\bar{\Omega}_{\cdot,j}\sqrt{n\log(m)} + O\big([n\bar{\Omega}_{\cdot,j}\log(m)]^{\frac{1}{2}}\big)}{n\bar{\Omega}_{\cdot,j}}$$

$$\le O\Big(\sqrt{\tfrac{\log(m)}{n}}\Big) + O\Big(\sqrt{\tfrac{\log(m)}{n\Omega_{\cdot,j}}}\Big). \tag{D.20}$$

This gives the second claim of Theorem C.1.

Next, we prove Theorem C.2. By (D.18) and (D.20), with probability $1-O(m^{-1})$, simultaneously for all $1 \le j \le m$,

$$|f_j - 1/2| \begin{cases} \ge \frac{\theta|T_j|}{2F_j} + O(e_n), & j \in S, \\ \le O(e_n), & j \in N, \end{cases}$$

where $e_n^2 = (\min\{1,\, \bar{s}\min_{j\in S} F_j,\, \min_{j\in N}\bar{\Omega}_{\cdot,j})^{-1}\frac{\log(m)}{n}$. The assumption (C.11) ensures that $\frac{\theta|T_j|}{2F_j} \gg e_n\sqrt{\log(m)}$. By setting the threshold at $e_n\sqrt{\log(\log(m))}$, all words in S will retain and all words in N will be screened out. □

D.2 Proof of Theorem C.3

Proof. By Theorem C.2, $\mathbb{P}(\widehat{S} = S) = 1 - o(1)$. Hence, we assume $\widehat{S} = S$ without loss of generality. In Algorithm 2, $\widehat{O}$ is obtained by modifying and renormalizing $\widetilde{O} = \widehat{D}\widehat{W}'(\widehat{W}\widehat{W}')^{-1}$. Since $\mathbb{E}\widehat{D} = OW$, we define a counterpart of $\widehat{O}$ by

$$O^* = OW\widehat{W}(\widehat{W}\widehat{W}')^{-1}.$$

Let $F^* = \frac{1}{2}(O^*_+ + O^*_-)$ and $T^* = \frac{1}{2}(O^*_+ - O^*_-)$. In the first part of our proof, we show that

$$\|F^* - F\|_1 = O(n^{-1}), \qquad \|T^* - \rho T\|_1 = O(n^{-1}) \tag{D.21}$$

In the second part of our proof, we show that

$$\|\widehat{O}_\pm - O^*_\pm\|_1 \le C\sqrt{|S|\log(m)/(n\bar{s})}. \tag{D.22}$$

The claim follows by combining (D.21)-(D.22).

First, we show (D.21). By definition,

$$\begin{aligned} [F^*, T^*] = O^* \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} &= O(W\widehat{W})(\widehat{W}\widehat{W}')^{-1} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \\ &= [F, T] \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} (W\widehat{W})(\widehat{W}\widehat{W}')^{-1} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}}_{\equiv M}. \end{aligned} \tag{D.23}$$

We now calculate the 2×2 matrix M. With the returns sorted in the ascending order, $y_{(1)} < y_{(2)} < \ldots < y_{(n)}$, Algorithm 2 sets $\widehat{p}_{(i)} = i/n$, for $1 \le i \le n$. It follows that

$$\widehat{W}\widehat{W}' = \begin{bmatrix} \sum_{i=1}^n \widehat{p}_i^2 & \sum_{i=1}^n (1-\widehat{p}_i)\widehat{p}_i \\ \sum_{i=1}^n (1-\widehat{p}_i)\widehat{p}_i & \sum_{i=1}^n (1-\widehat{p}_i)^2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n \widehat{p}_{(i)}^2 & \sum_{i=1}^n (1-\widehat{p}_{(i)})\widehat{p}_{(i)} \\ \sum_{i=1}^n (1-\widehat{p}_{(i)})\widehat{p}_{(i)} & \sum_{i=1}^n (1-\widehat{p}_{(i)})^2 \end{bmatrix}.$$

It is known that $\sum_{i=1}^n i = \frac{n(n+1)}{2}$ and $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$. We thereby calculate each entry of $\widehat{W}\widehat{W}'$: First, $\sum_{i=1}^n \widehat{p}_{(i)}^2 = \frac{1}{n^2}\sum_{i=1}^n i^2 = \frac{n}{3}[1+O(n^{-1})]$. Second, $\sum_{i=1}^n (1-\widehat{p}_{(i)})\widehat{p}_{(i)} = \frac{1}{n^2}\sum_{i=1}^n i(n-i) = \frac{1}{n}\sum_{i=1}^n i - \frac{1}{n^2}\sum_{i=1}^n i^2 = \frac{n}{6}[1+O(n^{-1})]$. Third, $\sum_{i=1}^n (1-\widehat{p}_{(i)})^2 = \frac{1}{n^2}\sum_{i=1}^n (n-i)^2 = \frac{1}{n^2}\sum_{i=0}^{n-1} i^2 = \frac{n}{3}[1+O(n^{-1})]$. Combining them gives

$$n^{-1}(\widehat{W}\widehat{W}') = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{bmatrix} + O(n^{-1}) \quad \Longrightarrow \quad n(\widehat{W}\widehat{W}')^{-1} = \begin{bmatrix} 4 & -2 \\ -2 & 4 \end{bmatrix} + O(n^{-1}). \tag{D.24}$$

Additionally, by direct calculations,

$$n^{-1}(W\widehat{W}') = \begin{bmatrix} \frac{1}{n}\sum_i p_i \widehat{p}_i & \frac{1}{n}\sum_i p_i(1-\widehat{p}_i) \\ \frac{1}{n}\sum_i (1-p_i)\widehat{p}_i & \frac{1}{n}\sum_i (1-p_i)(1-\widehat{p}_i) \end{bmatrix}. \tag{D.25}$$

We now plug (D.24)-(D.25) into (D.23). It gives

$$\begin{aligned} M &= \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{n}\sum_i p_i \widehat{p}_i & \frac{1}{n}\sum_i p_i(1-\widehat{p}_i) \\ \frac{1}{n}\sum_i (1-p_i)\widehat{p}_i & \frac{1}{n}\sum_i (1-p_i)(1-\widehat{p}_i) \end{bmatrix} \begin{bmatrix} 4 & -2 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \\ &= \begin{bmatrix} 1 & \frac{6}{n}\sum_i (\widehat{p}_i - \frac{1}{2}) \\ \frac{2}{n}\sum_i (p_i - \frac{1}{2}) & \frac{12}{n}\sum_i (p_i - \frac{1}{2})(\widehat{p}_i - \frac{1}{2}) \end{bmatrix}. \end{aligned}$$

The condition (C.8) yields $M_{21} = 0$. The way we construct $\{\widehat{p}_i\}_{i=1}^n$ ensures $M_{12} = O(n^{-1})$. Combined with the definition of ρ in (10), the above imply

$$M = \begin{bmatrix} 1 & 0 \\ 0 & \rho \end{bmatrix} + O(n^{-1}). \tag{D.26}$$

Then, (D.21) follows from plugging in (D.26) into (D.23).

Second, we show (D.22). Let $\overline{O} = [\overline{O}_+, \overline{O}_-]$ be the matrix obtained from setting negative entries of $\widetilde{O}$ to zero. Algorithm 2 outputs $\widehat{O}_\pm = (1/\|\overline{O}_\pm\|_1)\overline{O}_\pm$. It follows that, for $j \in S$,

$$|\widehat{O}_{\pm,j} - O^*_{\pm,j}| \leq |\overline{O}_{\pm,j} - O^*_{\pm,j}| + |\overline{O}_{\pm,j}| \cdot \left| \frac{1}{\|\overline{O}_\pm\|_1} - 1 \right|.$$

Since $\|O^*_\pm\|_1 = 1$, we have $|\|\overline{O}_\pm\|_1^{-1} - 1| = \|\overline{O}_\pm\|_1^{-1} |\|\overline{O}_\pm\|_1 - \|O^*_\pm\|_1| \leq \|\overline{O}_\pm\|_1^{-1} \|\overline{O}_\pm - O^*_\pm\|_1$. Hence,

$$|\widehat{O}_{\pm,j} - O^*_{\pm,j}| \leq |\overline{O}_{\pm,j} - O^*_{\pm,j}| + \frac{|\overline{O}_{\pm,j}|}{\|\overline{O}_\pm\|_1} \|\overline{O}_\pm - O^*_\pm\|_1. \tag{D.27}$$

Summing over j on both sides gives $\|\widehat{O}_\pm - O^*_\pm\|_1 \leq 2\|\overline{O}_\pm - O^*_\pm\|_1$. Moreover, since $O^*_\pm$ are nonnegative vectors, truncating out negative entries in $\overline{O}_\pm$ always makes it closer to $O^*_\pm$. It implies $\|\widetilde{O}_\pm - O^*_\pm\|_1 \leq \|\overline{O}_\pm - O^*_\pm\|_1$. Combining the above gives

$$\|\widehat{O}_\pm - O^*_\pm\|_1 \leq 2\|\widetilde{O}_\pm - O^*_\pm\|_1. \tag{D.28}$$

Therefore, to show (D.22), it suffices to bound $\|\widetilde{O}_\pm - O^*_\pm\|_1$.

Let W be the matrix whose i-th column is $(p_i, 1-p_i)'$. Since we have assumed $\widehat{S} = S$, it holds that $\widehat{d}_i = \widetilde{d}_i = s_i^{-1} d_i$. By model (4), $s_i \widehat{d}_i \sim \text{Multinomial}(s_i, p_i O_+ + (1-p_i) O_-)$. It leads to $\mathbb{E}\widehat{d}_i = (OW)_i$. Write $Z = \widehat{D} - \mathbb{E}\widehat{D}$. Then, $\widehat{D} = OW + Z$ and

$$\widetilde{O} = (OW + Z)\widehat{W}'(\widehat{W}\widehat{W}')^{-1} = O^* + Z\widehat{W}'(\widehat{W}\widehat{W}')^{-1}.$$

Let z_i be the i-th column of Z, $1 \le i \le n$. Plugging in the form of $\widehat{W}$, we have

$$Z\widehat{W}'(\widehat{W}\widehat{W}')^{-1} = \left[\sum_{i=1}^n \widehat{p}_i z_i \quad \sum_{i=1}^n (1-\widehat{p}_i) z_i \right] (\widehat{W}\widehat{W}')^{-1}.$$

It follows that

$$\begin{aligned} \|\widetilde{O}_{\pm,j} - O^*_{\pm,j}\|_1 &\le \max\left\{ \left| \frac{1}{n}\sum_{i=1}^n \widehat{p}_i Z_{i,j} \right|, \ \left| \frac{1}{n}\sum_{i=1}^n (1-\widehat{p}_i) Z_{i,j} \right| \right\} \left\| n(\widehat{W}\widehat{W}')^{-1} \right\|_1 \\ &\le C \max\left\{ \left| \frac{1}{n}\sum_{i=1}^n \widehat{p}_i Z_{i,j} \right|, \ \left| \frac{1}{n}\sum_{i=1}^n (1-\widehat{p}_i) Z_{i,j} \right| \right\}, \end{aligned} \tag{D.29}$$

where in the last line we have used (D.25). We now bound $|\frac{1}{n}\sum_{i=1}^n \widehat{p}_i Z_{i,j}|$. The bound for $|\frac{1}{n}\sum_{i=1}^n (1-\widehat{p}_i) Z_{i,j}|$ can be obtained similarly, so the proof is omitted. Since $\{\widehat{p}_i\}_{i=1}^n$ are constructed from $\{y_i\}_{i=1}^n$, they are independent of $\{Z_{i,j}\}_{i=1}^n$ by our assumption (1). We thus condition on $\{\widehat{p}_i\}_{i=1}^n$. Let $\{b_{i,j,\ell}\}_{\ell=1}^{s_i}$ be a collection of *iid* Bernoulli variables with a success probability $[p_i O_{+,j} + (1-p_i) O_{-,j}]$. Then, $d_{i,j}$ has the same distribution as $\sum_{\ell=1}^{s_i} b_{i,j,\ell}$. It follows that $Z_{i,j} \overset{(d)}{=} \sum_{\ell=1}^{s_i} s_i^{-1}(b_{i,j,\ell} - \mathbb{E}b_{i,j,\ell})$. Hence,

$$\sum_{i=1}^n \widehat{p}_i Z_{i,j} = \sum_{i=1}^n \sum_{\ell=1}^{s_i} \widehat{p}_i s_i^{-1}(b_{i,j,\ell} - \mathbb{E}b_{i,j,\ell}).$$

Conditioning on $\{\widehat{p}_i\}_{i=1}^n$, the variables $\widehat{p}_i s_i^{-1}(b_{i,j,\ell} - \mathbb{E}b_{i,j,\ell})$ are mutually independent, upper bounded by $2s_{\min}^{-1} \le C\bar{s}^{-1}$, each with mean 0 and variance $\le \bar{s}^{-2}(O_{+,j} + O_{+,j}) = 2\bar{s}^{-2}F_j$. By the Bernstein's inequality, with probability $1 - O(m^{-2})$,

$$\left| \sum_{i=1}^n \widehat{p}_i Z_{i,j} \right| \le C\sqrt{n\bar{s}^{-1}F_j \log(m)} + C\bar{s}^{-1}\log(m) \le C\sqrt{n\bar{s}^{-1}F_j \log(m)}, \tag{D.30}$$

where the last line is due to $n\bar{s}F_j/\log(m) \to \infty$. The bound for $|\sum_{i=1}^n (1-\widehat{p}_i) Z_{i,j}|$ is similar. Plugging them into (D.29) gives

$$\|\widetilde{O}_{\pm,j} - O^*_{\pm,j}\|_1 \le C\frac{\sqrt{F_j \log(m)}}{\sqrt{n\bar{s}}}. \tag{D.31}$$

It follows from Cauchy-Schwarz inequality that

$$\|\widetilde{O}_\pm - O^*_\pm\|_1 \le C\sqrt{\frac{\log(m)}{n\bar{s}}} \sum_{j\in S} \sqrt{F_j} \le C\sqrt{\frac{\log(m)}{n\bar{s}}} \cdot |S|^{\frac{1}{2}} \Big(\sum_{j \in S} F_j \Big)^{\frac{1}{2}} \le C\sqrt{\frac{|S|\log(m)}{n\bar{s}}}.$$

This proves (D.22). The proof is now complete. □

D.3 Proof of Theorem C.4

Proof. By Theorem C.2, $\mathbb{P}(\widehat{S} = S) = 1 - o(1)$. Hence, we assume $\widehat{S} = S$ without loss of generality.

We need some preparation. First, by our assumption, $F_j + \eta T_j = pO_{+,j} + (1-p)O_{-,j} \geq c_1(O_{+,j} + O_{-,j}) = 2c_1 F_j$. Second, by (D.31) in the proof of Theorem C.3, $|\widehat{F}_j - F_j| \leq C\sqrt{F_j \log(m)/(n\bar{s})}$ and $|\widehat{T}_j - \rho T_j| \leq C\sqrt{F_j \log(m)/(n\bar{s})}$. Since $n\bar{s}F_j \gg \log(m)$, we immediately obtain $|\widehat{F}_j - F_j| = o(F_j)$. Third, the condition (C.11) guarantees $n\theta^2 \frac{\rho^2 T_j^2}{F_j^2} \geq \frac{\log^2(m)}{\bar{s}F_j}$. In other words, $\rho|T_j| \gg \sqrt{F_j \log(m)/(n\bar{s})}$. So, $|\widehat{T}_j - \rho T_j| \ll \rho|T_j|$. We summarize these results as follows: for any $j \in S$,

$$\frac{|F_j + \eta T_j|}{F_j} \geq 2c_1, \qquad \frac{\max\{|\widehat{F}_j - F_j|, |\widehat{T}_j - \rho T_j|\}}{F_j} \leq C\sqrt{\frac{\log(m)}{n\bar{s}F_j}}, \qquad \frac{|\widehat{T}_j - \rho T_j|}{\rho|T_j|} = o(1). \tag{D.32}$$

We now proceed to the proof. Let $\eta = 2p - 1$ and $\widehat{\eta} = 2\widehat{p} - 1$. Then,

$$|\widehat{p} - p^*| = \frac{1}{2}|\widehat{\eta} - \rho^{-1}\eta|. \tag{D.33}$$

It suffices to bound $|\widehat{\eta} - \rho^{-1}\eta|$. We first show that the claim holds on the event $|\widehat{\eta} - \rho^{-1}\eta| \leq c_1$. We then show that this event holds with probability $1 - o(1)$.

Suppose $|\widehat{\eta} - \rho^{-1}\eta| \leq c_1$. Let $\widehat{F} = \frac{1}{2}(\widehat{O}_+ + \widehat{O}_-)$ and $\widehat{T} = \frac{1}{2}(\widehat{O}_+ - \widehat{O}_-)$. Since $p(1-p) = (1-\eta^2)/4$ and $p\widehat{O}_{+,j} + (1-p)\widehat{O}_{-,j} = \widehat{F} + \eta\widehat{T}_j$, the penalized MLE (A.3) has an equivalent form:

$$\widehat{\eta} = \mathrm{argmax}_{\eta \in [-1,1]} \ell_\lambda(\eta), \quad \text{where } \ell_\lambda(\eta) \equiv s^{-1}\sum_{j \in S} d_j \log(\widehat{F}_j + \eta\widehat{T}_j) + \lambda \log(1-\eta) + \lambda\log(1+\eta).$$

It follows that $\ell_\lambda(\widehat{\eta}) \geq \ell_\lambda(\rho^{-1}\eta)$. Rearranging the terms gives

$$s^{-1}\sum_{j \in S} d_j \log\Bigl(1 + \frac{(\widehat{\eta} - \rho^{-1}\eta)\widehat{T}_j}{\widehat{F}_j + \rho^{-1}\eta\widehat{T}_j}\Bigr) + \lambda\log\Bigl(1 + \frac{\widehat{\eta} - \rho^{-1}\eta}{1 + \rho^{-1}\eta}\Bigr) + \lambda\log\Bigl(1 - \frac{\widehat{\eta} - \rho^{-1}\eta}{1 - \rho^{-1}\eta}\Bigr) \geq 0. \tag{D.34}$$

Note that $1 + \rho^{-1}\eta = 2p^* \geq 2c_1$. So, on the event $|\widehat{\eta} - \rho^{-1}\eta| \leq c_1$, $\bigl|\frac{\widehat{\eta} - \rho^{-1}\eta}{1 + \rho^{-1}\eta}\bigr| \leq \frac{1}{2}$. Following a similar argument, we have $\bigl|\frac{\widehat{\eta} - \rho^{-1}\eta}{1 - \rho^{-1}\eta}\bigr| \leq \frac{1}{2}$. Note that $\log(1 \pm x) \leq \pm x - \frac{x^2}{4}$ for $x \in [-\frac{1}{2}, \frac{1}{2}]$. It follows that

$$\begin{aligned}
& \log\Bigl(1 + \frac{\widehat{\eta} - \rho^{-1}\eta}{1 + \rho^{-1}\eta}\Bigr) + \log\Bigl(1 - \frac{\widehat{\eta} - \rho^{-1}\eta}{1 - \rho^{-1}\eta}\Bigr) \\
\leq\ & \frac{\widehat{\eta} - \rho^{-1}\eta}{1 + \rho^{-1}\eta} - \frac{(\widehat{\eta} - \rho^{-1}\eta)^2}{4(1 + \rho^{-1}\eta)^2} - \frac{\widehat{\eta} - \rho^{-1}\eta}{1 - \rho^{-1}\eta} - \frac{(\widehat{\eta} - \rho^{-1}\eta)^2}{4(1 - \rho^{-1}\eta)^2} \\
=\ & -(\widehat{\eta} - \rho^{-1}\eta)\frac{2\rho^{-1}\eta}{1 - \rho^{-2}\eta^2} - (\widehat{\eta} - \rho^{-1}\eta)^2\frac{1 + \rho^{-2}\eta^2}{2(1 - \rho^{-2}\eta^2)^2}.
\end{aligned} \tag{D.35}$$

Also, by (D.32), $\widehat{F}_j + \rho^{-1}\eta\widehat{T}_j \sim F_j + \eta T_j \geq 2c_1 F_j$ and $|\widehat{T}_j| \sim \rho|T_j|$. Hence, $\bigl|\frac{(\widehat{\eta} - \rho^{-1}\eta)\widehat{T}_j}{\widehat{F}_j + \rho^{-1}\eta\widehat{T}_j}\bigr| \leq |\widehat{\eta} - \rho^{-1}\eta| \cdot \frac{\rho}{2c_1}$, which is bounded by $\frac{1}{2}$ on the event $|\widehat{\eta}-\rho^{-1}\eta|\leq c_1$. Note that $\log(1+x)\leq x-\frac{x^2}{4}$ for $x\in[-\frac{1}{2},\frac{1}{2}]$. We thus have

$$\begin{aligned}&s^{-1}\sum_{j\in S}d_j\log\left(1+\frac{(\widehat{\eta}-\rho^{-1}\eta)\widehat{T}_j}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j}\right)\\ \leq\;&(\widehat{\eta}-\rho^{-1}\eta)\sum_{j\in S}\frac{s^{-1}d_j\widehat{T}_j}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j}-(\widehat{\eta}-\rho^{-1}\eta)^2\sum_{j\in S}\frac{s^{-1}d_j\widehat{T}_j^2}{4(\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j)^2}.\end{aligned}\tag{D.36}$$

We plug (D.35)-(D.36) into (D.34). It gives

$$(\widehat{\eta}-\rho^{-1}\eta)X_1-(\widehat{\eta}-\rho^{-1}\eta)^2X_2\geq 0,\qquad\Longrightarrow\qquad|\widehat{\eta}-\rho^{-1}\eta|\leq\frac{|X_1|}{X_2},\tag{D.37}$$

where

$$X_1=\sum_{j\in S}\frac{s^{-1}d_j\widehat{T}_j}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j}-\frac{2\lambda\rho^{-1}\eta}{1-\rho^{-2}\eta^2},\qquad X_2=\sum_{j\in S}\frac{s^{-1}d_j\widehat{T}_j^2}{4(\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j)^2}+\frac{\lambda(1+\rho^{-2}\eta^2)}{2(1-\rho^{-2}\eta^2)^2}.$$

Below, we give an upper bound for $|X_1|$ and a lower bound for X_2.

Consider X_1. Since $(\widehat{F},\widehat{T})$ are obtained from the training data, they are independent of d. We thus condition on $(\widehat{F},\widehat{T})$. Using (D.32), we can get

$$\begin{aligned}&\left|\sum_{j\in S}\frac{\widehat{T}_js^{-1}\mathbb{E}d_j}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j}\right|\\ \leq&\left|\sum_{j\in S}\frac{\rho T_js^{-1}\mathbb{E}d_j}{F_j+\eta T_j}\right|+\left|\sum_{j\in S}\frac{(\widehat{T}_j-\rho T_j)s^{-1}\mathbb{E}d_j}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j}\right|+\left|\sum_{j\in S}\rho T_js^{-1}\mathbb{E}d_j\left(\frac{1}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j}-\frac{1}{F_j+\eta T_j}\right)\right|\\ \leq&\left|\sum_{j\in S}\frac{\rho T_js^{-1}\mathbb{E}d_j}{F_j+\eta T_j}\right|+\sum_{j\in S}\frac{|\widehat{T}_j-\rho T_j|s^{-1}\mathbb{E}d_j}{2(F_j+\eta T_j)}+\sum_{j\in S}\rho|T_j|s^{-1}\mathbb{E}d_j\frac{|\widehat{F}_j-F_j|+\rho^{-1}\eta|\widehat{T}_j-\rho T_j|}{2(F_j+\eta T_j)}\\ =&\left|\rho\sum_{j\in S}T_j\right|+\frac{1}{2}\sum_{j\in S}|\widehat{T}_j-\rho T_j|+\frac{1}{2}\sum_{j\in S}\rho|T_j|\big(|\widehat{F}_j-F_j|+\rho^{-1}\eta|\widehat{T}_j-\rho T_j|\big)\\ \leq&0+C\|\widehat{F}-F\|_1+C\|\widehat{T}-\rho T\|_1\\ \leq&C\sqrt{\frac{|S|\log(m)}{n\bar{s}}},\end{aligned}$$

where the second last line is due to $\sum_{j\in S}O_{+,j}=\sum_{j\in S}O_{-,j}=1$ and the last line is by Theorem C.3. Moreover, since the covariance matrix of d_j is $s\cdot\mathrm{diag}(F+\eta T)-s(F+\eta T)(F+\eta T)'\preceq s\cdot\mathrm{diag}(F+\eta T)$, we have

$$\mathrm{Var}\left(\sum_{j\in S}\frac{\widehat{T}_js^{-1}d_j}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j}\right)\leq\sum_{j\in S}\frac{\widehat{T}_j^2s^{-2}\cdot s(F_j+\eta T_j)}{(\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j)^2}$$

$$\begin{aligned}&\le Cs^{-1}\sum_{j\in S}\frac{\rho^2T_j^2}{F_j}+Cs^{-1}\sum_{j\in S}\frac{(\widehat{T}_j-\rho T_j)^2}{F_j}\\&\le Cs^{-1}\rho^2\Theta+Cs^{-1}\frac{|S|\log(m)}{n\bar{s}},\end{aligned}$$

where we have used (D.32). Let $\{b_\ell\}_{\ell=1}^s$ be *iid* variables, where $b_\ell\sim\text{Multinomial}(1,F+\eta T)$. Then, d has the same distribution as $\sum_{\ell=1}^s b_\ell$. It follows that

$$\sum_{j\in S}\frac{\widehat{T}_js^{-1}d_j}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j}\overset{(d)}{=}\sum_{\ell=1}^s\xi_\ell,\qquad\text{with}\quad\xi_\ell\equiv\sum_{j\in S}\frac{b_{\ell,j}\widehat{T}_j}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j}.$$

Conditioning on $(\widehat{F},\widehat{T})$, $\{\xi_\ell\}_{\ell=1}^s$ are *iid* variables, with $|\xi_\ell|\le\rho(2sc_1)^{-1}\sum_{j\in S}|b_{\ell,j}|\le\rho(2sc_1)^{-1}$. Also, in the above, we have derived the bound for $|\sum_{\ell=1}^s\mathbb{E}\xi_\ell|$ and $\text{Var}(\sum_{\ell=1}^s\xi_\ell)$. We apply the Bernstein's inequality and find out that, for any $\epsilon\in(0,1)$, with probability $1-\epsilon$,

$$\begin{aligned}\left|\sum_{j\in S}\frac{\widehat{T}_js^{-1}d_j}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j}\right|&\le C\sqrt{\frac{|S|\log(m)}{n\bar{s}}}+C\rho\sqrt{\frac{\Theta\log(\epsilon^{-1})}{s}}+\frac{\rho\log(\epsilon^{-1})}{2c_1s}\\&\le C\sqrt{\frac{|S|\log(m)}{n\bar{s}}}+C\rho\sqrt{\frac{\Theta\log(\epsilon^{-1})}{s}},\end{aligned}\tag{D.38}$$

where the last line is because $s\Theta\to\infty$. We plug (D.38) into the expression of X_1. Additionally, we notice that $1-\rho^{-2}\eta^{-2}=(1+\rho^{-1}\eta)(1-\rho^{-1}\eta)=4p^*(1-p^*)\ge4c_1^2$. Hence, with probability $1-\epsilon$,

$$|X_1|\le\frac{\lambda}{2c_1^2}|\rho^{-1}\eta|+C\sqrt{\frac{|S|\log(m)}{n\bar{s}}}+C\rho\sqrt{\frac{\Theta\log(\epsilon^{-1})}{s}}.\tag{D.39}$$

Consider X_2. It is seen that, conditioning on $(\widehat{F},\widehat{T})$,

$$\sum_{j\in S}\frac{\widehat{T}_j^2s^{-1}\mathbb{E}d_j}{4(\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j)^2}=\sum_{j\in S}\frac{\widehat{T}_j^2s^{-1}[s(F_j+\eta T_j)]}{4(\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j)^2}\ge C^{-1}\sum_{j\in S}\frac{\rho^2T_j^2}{F_j}\ge C^{-1}\rho^2\Theta.$$

At the same time,

$$\text{Var}\left(\sum_{j\in S}\frac{\widehat{T}_j^2s^{-1}d_j}{4(\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j)^2}\right)\le\sum_{j\in S}\frac{\widehat{T}_j^4s^{-2}[s(F_j+\eta T_j)]}{16(\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j)^4}\le Cs^{-1}\sum_{j\in S}\frac{\rho^4T_j^4}{F_j^3}\le Cs^{-1}\rho^4\Theta.$$

Similarly as proving (D.38), we then introduce variables $\{b_\ell\}_{\ell=1}^s$ and apply the Bernstein's inequality. Note that the above variance is much smaller than the square of the mean, due to $s\Theta\to\infty$. It follows that, with probability $1-\epsilon$,

$$\sum_{j\in S}\frac{\widehat{T}_j^2s^{-1}d_j}{4(\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j)^2}\ge C^{-1}\rho^2\Theta.\tag{D.40}$$

We plug (D.40) into the expression of X_2 and note that $1-\rho^{-2}\eta^2 = 4p^*(1-p^*) \leq 1$. It yields that

$$X_2 \geq \frac{\lambda}{2} + C^{-1}\rho^2\Theta. \tag{D.41}$$

We now plug (D.39) and (D.41) into (D.37). It follows that

$$|\widehat{\eta} - \rho^{-1}\eta| \leq C\,\frac{\lambda|\rho^{-1}\eta| + \sqrt{\frac{|S|\log(m)}{n\bar{s}}} + \rho\sqrt{\frac{\Theta\log(\epsilon^{-1})}{s}}}{\lambda + \rho^2\Theta}$$

By separating two cases, $\lambda \leq \rho^2\Theta$ and $\lambda > \rho^2\Theta$, we immediately obtain

$$|\widehat{\eta} - \rho^{-1}\eta| \leq C\begin{cases} \frac{\lambda}{\rho^2\Theta}|\rho^{-1}\eta| + \Big(\frac{\sqrt{|S|\log(m)}}{\rho^2\Theta\sqrt{n\bar{s}}} + \frac{\sqrt{\log(\epsilon^{-1})}}{\rho\sqrt{\Theta s}}\Big), & \text{if } \lambda \leq \rho^2\Theta, \\ |\rho^{-1}\eta| + \frac{\rho^2\Theta}{\lambda}\Big(\frac{\sqrt{|S|\log(m)}}{\rho^2\Theta\sqrt{n\bar{s}}} + \frac{\sqrt{\log(\epsilon^{-1})}}{\rho\sqrt{\Theta s}}\Big), & \text{if } \lambda > \rho^2\Theta. \end{cases}$$

Combining it with (D.33) and noting that $\rho^{-1}\eta = 2(p^* - \frac{1}{2})$, we have the desired claim.

What remains is to show that the event $|\widehat{\eta} - \rho^{-1}\eta| \leq c_1$ holds with probability $1 - o(1)$. For the function $\ell_\lambda(\cdot)$, by direct calculations,

$$\ell'_\lambda(\eta) = \sum_{j\in S}\frac{d_j\widehat{T}_j}{\widehat{F}_j + \eta\widehat{T}_j} - \frac{2\lambda\eta}{1-\eta^2}, \qquad \ell''_\lambda(\eta) = -\sum_{j\in S}\frac{d_j\widehat{T}_j^2}{2(\widehat{F}_j + \eta\widehat{T}_j)^2} - \frac{\lambda(1+\eta^2)}{2(1-\eta^2)^2}.$$

As $\eta \to +1$, $\ell'_\lambda(\eta) \to -\infty$; as $\eta \to -1$, $\ell'_\lambda(\eta) \to +\infty$. Hence, the maximum is attained in the interior of $(-1,1)$. Since the true $p^* \in [c_1, 1-c_1]$, it follows that $|\rho^{-1}\eta| \leq |1-2c_1|$. We now evaluate $\ell'_\lambda(\cdot)$ at $1-1.9c_1$. Following the same argument as proving (D.38), we can show that

$$\begin{aligned} \ell'_\lambda(1-2c_1) &= \sum_{j\in S}\frac{\rho T_j(F_j+\eta T_j)}{F_j + (1-1.9c_1)\rho T_j} - \frac{2\lambda(1-1.9c_1)}{[1-(1-1.9c_1)^2]^2} + O\Big(\sqrt{\tfrac{|S|\log(m)}{n\bar{s}}}\Big) + O\Big(\rho\sqrt{\tfrac{\Theta\log(\epsilon^{-1})}{s}}\Big) \\ &= -\sum_{j\in S}\frac{\rho^2[(1-1.9c_1) - \rho^{-1}\eta]T_j^2}{[F_j + (1-1.9c_1)\rho T_j]} - \frac{2\lambda(1-1.9c_1)}{[1-(1-1.9c_1)^2]^2} + o(\lambda + \rho^2\Theta) \\ &\geq -0.1c_1\rho^2\Theta - \frac{2\lambda(1-1.9c_1)}{[1-(1-1.9c_1)^2]^2} + O\Big(\sqrt{\tfrac{|S|\log(m)}{n\bar{s}}}\Big) + o(\lambda+\rho^2\Theta). \end{aligned}$$

So, it is strictly negative. As a result, the maximum cannot be attained at $[1-1.9c_1, 1)$. Similarly, we can prove that the maximum cannot be attained at $(-1, -1+1.9c_1]$. Now, we have restricted our attention to a compact interval that is bounded away from ± 1 by at least $1.9c_1$. For any η_0 in this interval, $F_j + \eta_0 T_j \geq cF_j$ for a constant $c > 0$. This allows us to mimic the proof of (D.40)-(D.41) to get

$$-\ell''_\lambda(\eta_0) \geq C^{-1}(\lambda + \rho^2\Theta), \qquad \text{for } \eta_0 \text{ in this compact interval.}$$

By Taylor expansion, there exists η_0, whose value is between $\rho^{-1}\eta$ and $\widehat{\eta}$, such that

$$0 = \ell'_\lambda(\widehat{\eta}) = \ell'_\lambda(\rho^{-1}\eta) + \ell''_\lambda(\eta_0)(\widehat{\eta} - \rho^{-1}\eta).$$

If $|\widehat{\eta} - \rho^{-1}\eta| > c_1$, then the above implies $|\ell'_\lambda(\rho^{-1}\eta)| \geq c_1|\ell''_\lambda(\eta_0)| \geq C^{-1}(\lambda + \rho^2\Theta)$. On the other hand, we notice that $X_1 = \ell'_\lambda(\rho^{-1}\eta)$, where we have proved in (D.39) that $|X_1| = o(\lambda + \rho^2\Theta)$. This yields a contradiction. The proof is now complete. □

D.4 Proof of Theorem C.5

Proof. Since $\{p_i\}_{i=1}^N$ are drawn from a continuous density, with probability 1, their values are distinct from each other. The Spearman correlation coefficient has an equivalent form:

$$SR(\hat{p}, p) = 1 - \frac{1}{N(N^2-1)} \sum_{i=1}^N (\hat{r}_i - r_i)^2, \tag{D.42}$$

where r_i is the rank of p_i among $\{p_i\}_{i=1}^N$, which also equals to the rank of p_i^* among $\{p_i^*\}_{i=1}^N$, and $\hat{r}_i$ is the rank of $\hat{p}_i$ among $\{\hat{p}_i\}_{i=1}^N$. By definition,

$$r_i = \sum_{j=1}^N \mathrm{sgn}(p_i^* - p_j^*) + N + 1, \qquad \hat{r}_i = \sum_{j=1}^N \mathrm{sgn}(\hat{p}_i - \hat{p}_j) + N + 1,$$

where the sign function takes values in $\{0, \pm 1\}$. In the proof of Theorem C.4, letting $\epsilon = N^{-2}$, we get the following result: Conditioning on $\{p_i\}_{i=1}^N$, with probability $1 - N^{-2}$,

$$\max_{1 \leq i \leq N} |\hat{p}_i - p_i^*| \leq \delta, \qquad \text{where} \quad \delta = \frac{C}{\rho\sqrt{\Theta}} \left(\frac{\sqrt{|S|\log(m)}}{\rho\sqrt{n\bar{s}\Theta}} + \frac{\sqrt{\log(N)}}{\sqrt{s}} \right). \tag{D.43}$$

We note that the quantity ρ on the right hand side depends on the training labels while the probability law is with respect to the randomness of the training and testing articles. By the assumption (1), we can always condition on the training labels and treat ρ as a constant. Let D be the event that (D.43) holds simultaneously for all $1 \leq i \leq N$. Using the probability union bound, we have $\mathbb{P}(D) = 1 - N^{-1}$. For each $1 \leq i \leq N$, define the index set

$$B_i(3\delta) = \{1 \leq j \leq N : j \neq i,\ |p_j^* - p_i^*| \leq 3\delta\}.$$

On the event D, for $j \notin B_i(3\delta)$, $|p_i^* - p_j^*| > 3\delta$, while $|\hat{p}_i - p_i^*| \leq \delta$ and $|\hat{p}_j - p_j^*| \leq \delta$; hence, $(\hat{p}_i - \hat{p}_j)$ must have the same sign as $(p_i^* - p_j^*)$. It follows that

$$|r_i - r_j| \leq \sum_{j \in B_i(3\delta)} \big(|\mathrm{sgn}(p_i^* - p_j^*)| + |\mathrm{sgn}(\hat{p}_i - \hat{p}_j)| \big) \leq 2|B_i(3\delta)|.$$

We plug it into (D.42) and note that $|\hat{r}_i - r_i|^2 \leq N|\hat{r}_i - r_i|$. It yields

$$1 - SR(\hat{p}, p) \leq \frac{1}{N^2 - 1} \sum_{i=1}^N |\hat{r}_i - r_i| \leq \frac{2N}{N^2-1} \max_{1 \leq i \leq N} |B_i(3\delta)|. \tag{D.44}$$

In other words, conditioning on $\{p_i^*\}_{i=1}^N$, (D.44) holds with probability $1 - N^{-1}$.

We now bound $|B_i(3\delta)|$, taking into consideration the randomness of $\{p_i^*\}_{i=1}^N$. Each p_i^* is a non-random, linear, monotonically increasing function of p_i (note: ρ is treated as non-random; see explanations above). Therefore, the distribution assumption on $\{p_i\}_{i=1}^N$ yields that $\{p_i^*\}_{i=1}^N$ are *iid* drawn from a continuous distribution on $[c_1, 1-c_1]$. The probability density of this distribution must be Lipschitz. Fix $1 \le i \le N$ and write

$$|B_i(3\delta)| = \sum_{j\neq i} 1\Big\{p_j^* \in [p_i^* - 3\delta, p_i^* + 3\delta] \cap [c_1, 1-c_1]\Big\}.$$

Conditioning on p_i^*, the other p_j^*'s are *iid* drawn from a Lipschitz probability density. As a result, each other p_j^* has a probability of $O(\delta)$ to fall within a distance of 3δ to p_i^*, i.e., $|B_i(3\delta)|$ is the sum of $(N-1)$ *iid* Bernoulli variables with a success probability of $O(\delta)$. By the Bernstein's inequality, with probability $1 - N^{-2}$,

$$|B_i(3\delta)| \le CN\delta + C\sqrt{N\delta\log(N)} + C\log(N).$$

Combining it with the probability union bound, with probability $1 - N^{-1}$, the above inequality holds simultaneously for all $1 \le i \le N$. We then plug it into (D.44) and get

$$1 - SR(\hat{p}, p) \le C\delta + C\sqrt{\frac{\delta\log(N)}{N}} + \frac{C\log(N)}{N} \le C\max\Big\{\delta,\ \frac{\log(N)}{N}\Big\}. \tag{D.45}$$

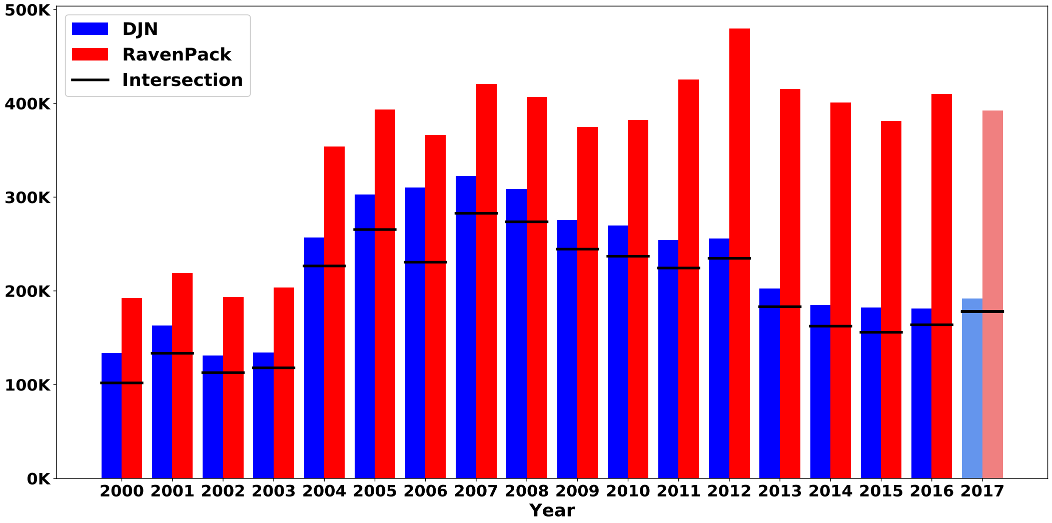
Under our assumption, the right hand side of (D.45) is $o(1)$. The claim follows immediately. □

E RavenPack

The data we use are composite sentiment scores from RavenPack News Analytics 4 (RPNA4) DJ Edition Equities. The underlying news data for this version of RavenPack should be identical to the collection of Dow Jones articles that we use to build SSESTM. However, the observation count that we see in RavenPack is somewhat larger than the number of observations we can construct from the underlying Dow Jones news. The discrepancy arises from the black-box transformations that RavenPack applies during it analytics process. Ultimately, what we observe in RavenPack is their collection of article-level score that is indexed by stock ticker and time, and it is not possible to accurately map RavenPack observations back to the original news. As a result, we cannot pin down the precise source of the difference in observation counts between our two data sets. The most likely explanation is that RavenPack uses a proprietary algorithm to assign ticker tags to articles, while we rely on the tags assigned directly by Dow Jones. If this expands the set of articles tagged to a given firmTo the extent that RavenPack adds tags, it is likely that the same article is assigned to multiple firms, which results in a higher RavenPack observation count.

Figure A.3 shows the differences in observation counts in our data set (the complete set of Dow Jones Newswires from 1984 through mid-2017) versus RavenPack. We restrict all counts to those having a uniquely matched stock identifier in CRSP. We see that early in the sample the article

Figure A.3: Dow Jones Newswire and RavenPack Observation Counts



count for Newswires and RavenPack are similar, but this difference grows over time. When we map Newswires to CRSP, we use articles' stock identifier tags, which are provided by Dow Jones. Our interpretation of the figure is that, over time, RavenPack has become more active in assigning their own stock assignments to previously untagged articles.

F Additional Exhibits

Table A.2: List of Top 50 Positive/Negative Sentiment Words

Positive			Negative		
Word	Score	Samples	Word	Score	Samples
repurchase	0.573	14	shortfall	0.323	14
surpass	0.554	14	downgrade	0.382	14
upgrade	0.551	14	disappointing	0.392	14
undervalue	0.604	13	tumble	0.402	14
surge	0.551	13	blame	0.414	14
customary	0.549	11	hurt	0.414	14
jump	0.548	11	auditor	0.424	14
declare	0.545	11	plunge	0.429	14
rally	0.568	10	slowdown	0.433	14
discretion	0.544	10	plummet	0.418	13
beat	0.538	10	miss	0.424	13
treasury	0.567	9	waiver	0.418	12
unsolicited	0.555	9	sluggish	0.428	12
buy	0.548	9	downward	0.433	12
climb	0.543	9	warn	0.435	12
tender	0.541	9	halt	0.417	11
top	0.540	9	lower	0.424	11
imbalance	0.567	8	fall	0.431	11
up	0.568	7	resign	0.441	11
bullish	0.555	7	soften	0.443	11
soar	0.548	7	slash	0.435	10
tanker	0.546	7	lackluster	0.437	10
deepwater	0.544	7	postpone	0.445	10
reconnaissance	0.544	7	unfortunately	0.445	10
fastener	0.538	7	unlawful	0.447	10
bracket	0.538	7	covenant	0.424	9
exceed	0.534	7	woe	0.425	9
visible	0.557	6	delay	0.428	9
valve	0.545	6	subpoena	0.429	9
unanimously	0.543	6	default	0.437	9
bidder	0.540	6	soft	0.437	9
terrain	0.539	6	widen	0.438	9
gratify	0.536	6	issuable	0.441	9
armor	0.536	6	regain	0.441	9
unregistered	0.535	6	deficit	0.442	9
tag	0.559	5	irregularity	0.442	9
maritime	0.542	5	bondholder	0.445	9
reit	0.542	5	weak	0.445	9
warfare	0.539	5	hamper	0.445	9
propane	0.539	5	notify	0.451	9
hydraulic	0.534	5	insufficient	0.433	8
epidemic	0.534	5	unfavorable	0.434	8
horizon	0.582	4	erosion	0.436	8
clip	0.567	4	allotment	0.446	8
potent	0.566	4	suspend	0.454	8
fragment	0.562	4	inefficiency	0.434	7
fossil	0.550	4	persistent	0.435	7
reallowance	0.549	4	worse	0.439	7
terrorism	0.544	4	setback	0.443	7
suburban	0.539	4	parentheses	0.445	7

Note: The table shows the list of top 50 lists words with positive and negative sentiment based on screening from the 14 training and validation samples. These 50 words are selected by first sorting on the number of samples (out of 14) in which the word was selected, and then sorting on their average sentiment score.