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

October 10, 2020

Abstract

We introduce a new text-mining methodology that extracts 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 score that is specifically adapted to the problem of return prediction. Our method proceeds in three steps: 1) isolating a list of terms via predictive screening, 2) assigning prediction weights to these words via topic modeling, and 3) aggregating terms into an article-level predictive score via penalized likelihood. We derive theoretical guarantees on the accuracy of estimates from our model with minimal assumptions. In our empirical analysis, we study one of the most actively monitored streams of news articles in the financial system—the *Dow Jones Newswires*—and show that our supervised text model excels at extracting return-predictive signals in this context. Information in newswires is assimilated into prices with an inefficient delay that is broadly consistent with limits-to-arbitrage (i.e., more severe for smaller and more volatile firms) yet can be exploited in a real-time trading strategy with reasonable turnover and net of transaction costs.

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 Torben Andersen, Frank Diebold, Robert Engle, Timothy Loughran, Xavier Gabaix, as well as seminar and conference participants at the New York University, Yale University, University of Pennsylvania, University of Southern California, Cheung Kong Graduate School of Business, Ohio State University, Hong Kong University of Science and Technology, University of Zurich, Peking University, University of Liverpool, Zhejiang University, AQR, T. Rowe Price, NBER Conference on Big Data: Long-Term Implications for Financial Markets and Firms, the 12th SoFiE Annual Meeting, Canadian Econometrics Study Group Meeting, China International Conference in Finance, Market Microstructure and High Frequency Data Conference at the University of Chicago, the Panel Data Forecasting Conference at USC Dornsife Institute, JHU Carey Finance Conference, SIAM conference on Financial Mathematics & Engineering, SAIF International Conference on FinTech, FinTech Symposium in Guanghua School of Management, and 3rd Chinese Econometricians Forum. 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 practical 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 is most commonly used in finance to study the "sentiment" of a given document, and this sentiment is 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 phenomena such as information transmission in financial markets (Tetlock, 2014).

In this paper we present a novel model-based approach to understanding the sentimental structure of a text corpus without relying on pre-existing dictionaries. Our model is motivated by the view that return-predictive content of a given event is reflected both in the news article text and in the returns of related assets. We propose a model that describes the *joint* generation of text and returns, where the document-level sentiment is represented by a latent variable. Our model is characterized by as few as two simple equations, which makes it a very flexible natural language processing (NLP) module that can be incorporated into more complex problems of a similar nature. In our model, the sentiment dictionary and term weights are parameters, and they can be estimated using the training data from a past study. Therefore, our approach allows for the construction of a sentiment dictionary that is customized for the context of interest and the data at hand. An important distinction of our approach with the literature is that our method will extract general *return predictive* content of news. The method does not differentiate between non-sentiment (i.e., objective information) and sentiment content of news per se. Nonetheless, in alignment with preceding literature, we continue to use the term "sentiment" to refer to the predictive signal that we extract from news.

We abbreviate our method as SESTM (Sentiment Extraction via Screening and Topic Modeling), and it consists of three parts. The first step isolates the most relevant terms from a very large vocabulary of terms via predictive correlation screening. The second step assigns term-specific sentiment weights using a supervised topic model. The third step uses the estimated topic model to assign article-level sentiment scores via penalized maximum likelihood.

The method we introduce has three main virtues. The first is simplicity—it requires only standard econometric techniques like correlation analysis and maximum likelihood estimation. Unlike commercial platforms or deep learning approaches which amount to black boxes for their users, the supervised learning approach we propose is entirely "white box." Second, our method requires mini-

mal 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 data set 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*. Ours is the first paper (to our knowledge) to analyze this data set, which is a central information source for broad swathes of market participants and is widely subscribed and closely monitored by sophisticated investors. It is 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 data from CRSP in order to model return behavior as a function of a Newswire content. The key feature of our approach is that we learn the sentiment scoring model from the joint behavior of article text and stock returns, rather than taking sentiment scores off the shelf.

To translates the statistical gains of our model into economic terms, 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). It does so by isolating an interpretable and intuitive ranking of positive and negative sentiment values for words in our corpus.

We compare the price impact of "fresh" versus "stale" news by devising a measure of article novelty. Stale articles are defined as those bearing close similarity to articles about the same stock over the preceding week. While the sentiment of stale news has a weakly significant positive association with future price changes, the effect 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 how differences in news assimilation associate with a variety of stock attributes. We find that price responses to news are roughly four times as large for smaller stocks (below NYSE median) and more volatile stocks (above median), and that it takes roughly twice as long for news about small and volatile stocks to be fully reflected in prices.

We establish a number of theoretical results to accompany our model. First is our theoretical guarantee of exactly recovering the sentiment dictionary from training data via a correlation screening step. It is reminiscent of the theory for marginal screening in regression models (Fan and Lv, 2008), but our model is very different as it tackles with count data. Second, we derive sharp error bounds for parameter estimation. The error bounds depend on the scale of the corpus (e.g., size of the vocabulary, total number of text documents, average number of words per document, etc.) and the

strength of sentiment signals (e.g., sensitivity of returns to sentiment, sensitivity of text generation to sentiment, etc.). Third, we derive and quantify the error of predicting the sentiment score of a newly arriving article.

Our paper contributes to a growing literature on methods for integrating textual analysis into empirical economics research (surveyed in Gentzkow et al., 2019). Most prior work using text as data for finance and accounting research does little direct statistical analysis of text, and instead relies on pre-defined sentiment dictionaries. Early examples are Tetlock (2007), who applies the Harvard-IV psychosocial dictionary to a subset of articles from *The Wall Street Journal*, and Loughran and McDonald (2011), who manually create a new sentiment dictionary specifically designed for the finance context. These papers manage the dimensionality challenge by restricting their analysis to words in pre-existing sentiment dictionaries and using ad hoc word-weighting schemes.

Some papers adopt off-the-shelf machine learning techniques, e.g., naïve Bayesian classifiers or support vector machines, to battle the curse of dimensionality such as Antweiler and Frank (2005), Li (2010), Manela and Moreira (2017) and Jegadeesh and Wu (2013). While potentially more flexible than basic dictionary methods, these approaches are implemented in ad hoc ways and the properties of estimators in these contexts are poorly understood. In contrast, we propose a text-based model and estimator that is specifically designed for the return prediction context. We rigorously prove the consistency of our estimator based precisely on how it is implemented (without ad hoc modifications). By showing how a financial objective can be integrated with rigorous analysis of sophisticated text models, we outline for a new research agenda for leveraging text data in both empirical finance and financial econometrics.

Furthermore, our empirical analysis is much more far-ranging than previous text-based analyses of stock returns. Most previous analysis focuses on text of either SEC filings or limited news such as the front page *The Wall Street Journal*. In contrast, we study *all* articles disseminated by the Dow Jones Newswires service since 1984, a data set whose breadth is unprecedented in text-based financial research. This includes *all* articles in *The Wall Street Journal* (not to mention press release wires, Barrons, MarketWatch, and the full range of Dow Jones realtime news services), covers a significantly longer sample than that available from the SEC (whose digital text is available only since 1993), and includes news content that arrives far more frequently (several thousand times per year for some firms) and with less severe scrubbing and self-reporting biases than SEC disclosures.

The closest benchmark for our analysis lies not in the academic literature, but comes from a commercial vendor of financial news sentiment scores. This firm, RavenPack (see Appendix D for details), is the natural benchmark for our study for two reasons. First, according to its marketing materials, it uses sophisticated (though proprietary and undisclosed) NLP methods to extract sentiment scores from financial news text. Thus RavenPack is closer to our work from a methodological standpoint than previous finance literature. Second, sentiment scores sold by RavenPack are derived from the exact same Dow Jones Newswires data set we study. Thus RavenPack is also the most appropriate empirical benchmark. Fortunately, through our subscription to RavenPack, we are able to make direct comparisons of our model versus its. We show in a head-to-head trading strategy analysis that our SESTM method translates into an equal-weighted portfolio Sharpe ratio

32% higher than that of RavenPack (17% higher for a value-weighted portfolio). And SESTM does so with complete model transparency, in contrast to the propriety block box of RavenPack.

The rest of the paper is organized as follows. In Section 2, we introduce a probabilistic model for sentiment analysis. In Section 3, we propose our SESTM method. Section 4 reports an empirical analysis of stock-level news and returns using SESTM. In Section 5, we describe the estimator's statistical properties, and we provide supporting mathematical proofs and Monte Carlo simulations in the appendix.

2 A Probabilistic Model for Sentiment Analysis

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 $d_{j,i}$ is the number of times word j occurs in article i. In matrix form, this is an $m \times n$ document-term matrix, $D = [d_1, ..., d_n]$. We occasionally work with a subset of rows from D, where the indices of columns included in the subset are listed in the set S. We denote the corresponding submatrix as $D_{[S],\cdot}$. We then use $d_{[S],i}$ to denote the column vector corresponding to the i^{th} column of $D_{[S],\cdot}$.

Articles are tagged with the identifiers of stocks mentioned in the articles. For simplicity, we study articles that correspond to a single stock,[1] and we label article i with the associated stock return (or its idiosyncratic component), y_i, on the publication date of the article.

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. The sentiment score p_i links the realized returns y_i with the word vector d_i, so we need at least these two 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. Given that our sole agenda is on sentiment, this is perhaps the simplest abstraction that suits our purpose.

To begin with, we wish to model the distribution of y_i given p_i as flexibly as possible in order to accommodate a wide range of potential associations between returns and sentiment. For the conditional return distribution, we assume

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

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

We now turn to the conditional distribution of word counts in an article. We model d_i by adapting the popular probabilistic topic model (Hofmann, 1999) to accommodate sentiment information. First,

[1] 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.

we assume the vocabulary has a partition:

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

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 vocabulary (S and N have dimensions $|S|$ and $m - |S|$, respectively). Likewise, $d_{[S],i}$ and $d_{[N],i}$ are the corresponding subvectors of d_i and contain counts of sentiment-charged and sentiment-neutral words, respectively. The distribution of sentiment-neutral counts, $d_{[N],i}$, is essentially a nuisance, so we leave it unmodeled and assume it is independent of the vector of interest, $d_{[S],i}$.

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

$$d_{[S],i} \sim \text{Multinomial}\Big(s_i,\ p_i O_+ + (1 - p_i) O_-\Big), \tag{3}$$

where s_i is the total count of sentiment-charged words in article i and therefore 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 word frequencies in a maximally positive sentiment article (one for which $p_i = 1$). Likewise, O_- is a "negative sentiment topic" that describes the distribution of word frequencies in maximally negative articles (those for which $p_i = 0$). At intermediate values of sentiment $0 < p_i < 1$, word frequencies are a convex combination of those from the positive and negative sentiment topics.

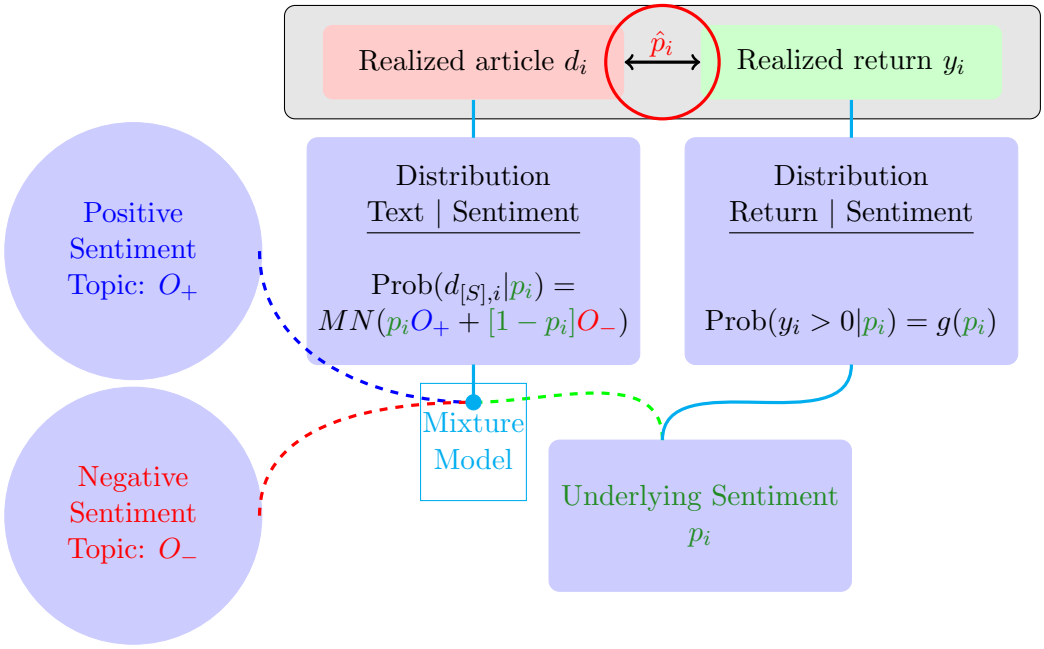
$O_\pm$ captures information on both the frequency of words as well as their sentiment. It is helpful, in fact, to reorganize the topic vectors into a *vector of frequency*, F, and a *vector of tone*, T:

$$F = \frac{1}{2}(O_+ + O_-), \qquad T = \frac{1}{2}(O_+ - O_-). \tag{4}$$

If a word has a larger value in F, it appears more frequently overall. But term-level sentiment is captured by the vector of tone. A word j has positive tone if it has a larger weight in the positive sentiment topic than in the negative sentiment topic; i.e., the j^{th} entry of T is positive (and likewise for negative tone). The absolute value of T captures the strength of tone.

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-charged 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_-.

Figure 1: Model Diagram



Note: Illustration of model structure.

3 SESTM: A Supervised Sentiment Extraction Algorithm

We now present our Sentiment Extraction via Screening and Topic Modeling (SESTM) estimation procedure, which consists of three steps. First, we screen for the set of sentiment-charged words. Second, we estimate the positive and negative sentiment topics O_+ and O_-. Third, we use penalized maximum likelihood to estimate sentiment scores of new articles. Sections 3.1 through 3.3 describe each step in detail.

3.1 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 vocabulary that accounts for the full joint distribution of sentiment-charged versus sentiment-neutral terms is at 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

$$f_j = \frac{\text{count of word } j \text{ in articles with sgn}(y) = +1}{\text{count of word } j \text{ in all articles}}. \tag{5}$$

for each $j = 1, ..., m$. If we view $\text{sgn}(y)$ as the response variable and the count of each word j as a predictor, then f_j can be viewed as a form of marginal screening statistics (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 compare f_j with proper thresholds. Let $\widehat{\pi}$ denote the fraction of articles tagged with a positive return in our training sample. For a sentiment neutral word, since its occurrence is uncorrelated with the sign of returns, we expect to see $f_j \approx \widehat{\pi}$. Hence, we set an upper threshold, α_+, and define all words having $f_j > \widehat{\pi} + \alpha_+$ as positive sentiment terms. Likewise, any word satisfying $f_j < \widehat{\pi} - \alpha_-$ for some lower threshold α_- is deemed a negative sentiment term. Finally, we impose a third threshold, κ, on the count of word j in all articles (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. Our estimate of the set S is defined by

$$\widehat{S} = \left\{ j : f_j \geq \widehat{\pi} + \alpha_+, \text{ or } f_j \leq \widehat{\pi} - \alpha_- \right\} \cap \{ j : k_j \geq \kappa \}. \tag{6}$$

The thresholds $(\alpha_+, \alpha_-, \kappa)$ are hyper-parameters that can be tuned via cross-validation.

Remark 1 (A variant of the screening step)**.** *We introduce a variant of the screening statistic that is particularly useful in empirical analysis:*

$$f_j^* = \frac{\textit{count of articles including word } j \textit{ AND having sgn}(y) = 1}{\textit{count of articles including word } j}. \tag{7}$$

It modifies f_j by truncating the count of word j in any individual article at the value 1*. While f_j may be sensitive to extreme values in article-specific word counts, e.g., a specific term mentioned many times in one article, f_j^* will not, and we find in empirical analyses that replacing f_j by f_j^* improves predictive performance.*

This procedure has similar theoretical properties as the screening procedure based on f, but the conditions based on f are more elegant and transparent, so we choose to present theory using f.

3.2 Learning Sentiment Topics

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.

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. In our setting, each Newswire is associated with a stock return, and the return contains information about article sentiment. Hence, returns serve as training labels and, in a low signal-to-noise ratio environment, there are likely to be efficiency gains from exploiting such labels via supervised learning. We therefore take a supervised learning approach to estimate O (or, equivalently, to estimate F and T) in the spirit of Mcauliffe and Blei (2008).

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 topic. Suppose, for now, that we observe these sentiment scores for all articles in our sample. Let $h_i = d_{[S],i}/s_i$ denote the $|S| \times 1$ vector of word frequencies. Model (3) implies that

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

or, in matrix form,

$$\mathbb{E}H = OW, \qquad \text{where} \quad W = \begin{bmatrix} p_1 & \cdots & p_n \\ 1-p_1 & \cdots & 1-p_n \end{bmatrix}, \quad \text{and} \quad H = [h_1, h_2, \ldots, h_n].$$

Based on this fact, we propose a simple approach to estimate O via a regression of H on W. Note that we do not directly observe H (because S is unobserved) or W. We estimate H by plugging in $\widehat{S}$ from the screening step:

$$\widehat{h}_i = d_{[\widehat{S}],i}/\widehat{s}_i, \qquad \text{where} \quad \widehat{s}_i = \sum_{j \in \widehat{S}} d_{j,i}. \tag{8}$$

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

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

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 several attractive features. First, it is simple to use. Second, it is sufficient to achieve statistical guarantees for our algorithm under weak assumptions. Third, it is robust to outliers that riddle the return data.

Given $\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_n$, we construct

$$\widehat{O} = \left[\widehat{h}_1, \widehat{h}_2, \ldots, \widehat{h}_n\right] \widehat{W}'(\widehat{W}\widehat{W}')^{-1}, \qquad \text{where} \quad \widehat{W} = \begin{bmatrix} \widehat{p}_1 & \widehat{p}_2 & \cdots & \widehat{p}_n \\ 1-\widehat{p}_1 & 1-\widehat{p}_2 & \cdots & 1-\widehat{p}_n \end{bmatrix}. \tag{10}$$

$\widehat{O}$ may have negative entries. We set all negative entries of this matrix to zero and re-normalize each column to have a unit ℓ^1-norm. To simplify notation, we reuse $\widehat{O}$ for the resulting matrix. We also use $\widehat{O}_\pm$ to denote the two columns of $\widehat{O} = [\widehat{O}_+, \widehat{O}_-]$.

3.3 Scoring New Articles

The preceding steps construct estimators $\widehat{S}$ and $\widehat{O}$. We now discuss how to estimate the sentiment p for a new article that is not included from the training sample. Let d be the the article's count vector and let s be the total count of sentiment-charged words. According to our model (3),

$$d_{[S]} \sim \text{Multinomial}\Big(s,\ pO_+ + (1-p)O_-\Big).$$

Given estimates $\widehat{S}$ and $\widehat{O}$, we can estimate p 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, $\lambda \log(p(1-p))$, in the likelihood function, and solve the following optimization:

$$\widehat{p} = \arg\max_{p\in[0,1]} \Big\{ \widehat{s}^{-1} \sum_{j\in\widehat{S}} d_j \log\Big(p\widehat{O}_{+,j} + (1-p)\widehat{O}_{-,j}\Big) + \lambda \log\big(p(1-p)\big) \Big\}, \tag{11}$$

where $\widehat{s}$ is the total count of words from $\widehat{S}$ in the new article, $(d_j, \widehat{O}_{+,j}, \widehat{O}_{-,j})$ are the jth entries of the corresponding vectors, and $\lambda > 0$ is a tuning parameter. 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 toward a neutral sentiment score of $1/2$, where the amount of shrinkage depends on the magnitude of λ.[2] 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.

To summarize: Step 1 estimates the sentiment-charged dictionary S from (5) and (6), Step 2 estimates the vectors of positive and negative sentiment scores, O_+ and O_-, from (9) and (10), and Step 3 predicts the sentiment score of a new article via (11).

4 Empirical Analysis

In this section, we apply our text-mining framework to the problem of return prediction for investment portfolio construction. First, our sentiment model uncovers strong predictive associations between news text and subsequent returns. Second, it translates the extent of predictability from statistical terms such as predictive R^2 into more meaningful economic terms, such as the growth rate in an investor's savings attributable to harnessing text-based information.

Our null hypothesis of market efficiency predicts that the expected return is dominated by unforecastable news, as this news is rapidly (in its starkest form, immediately) incorporated in prices. The maintained alternative 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

[2] 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.

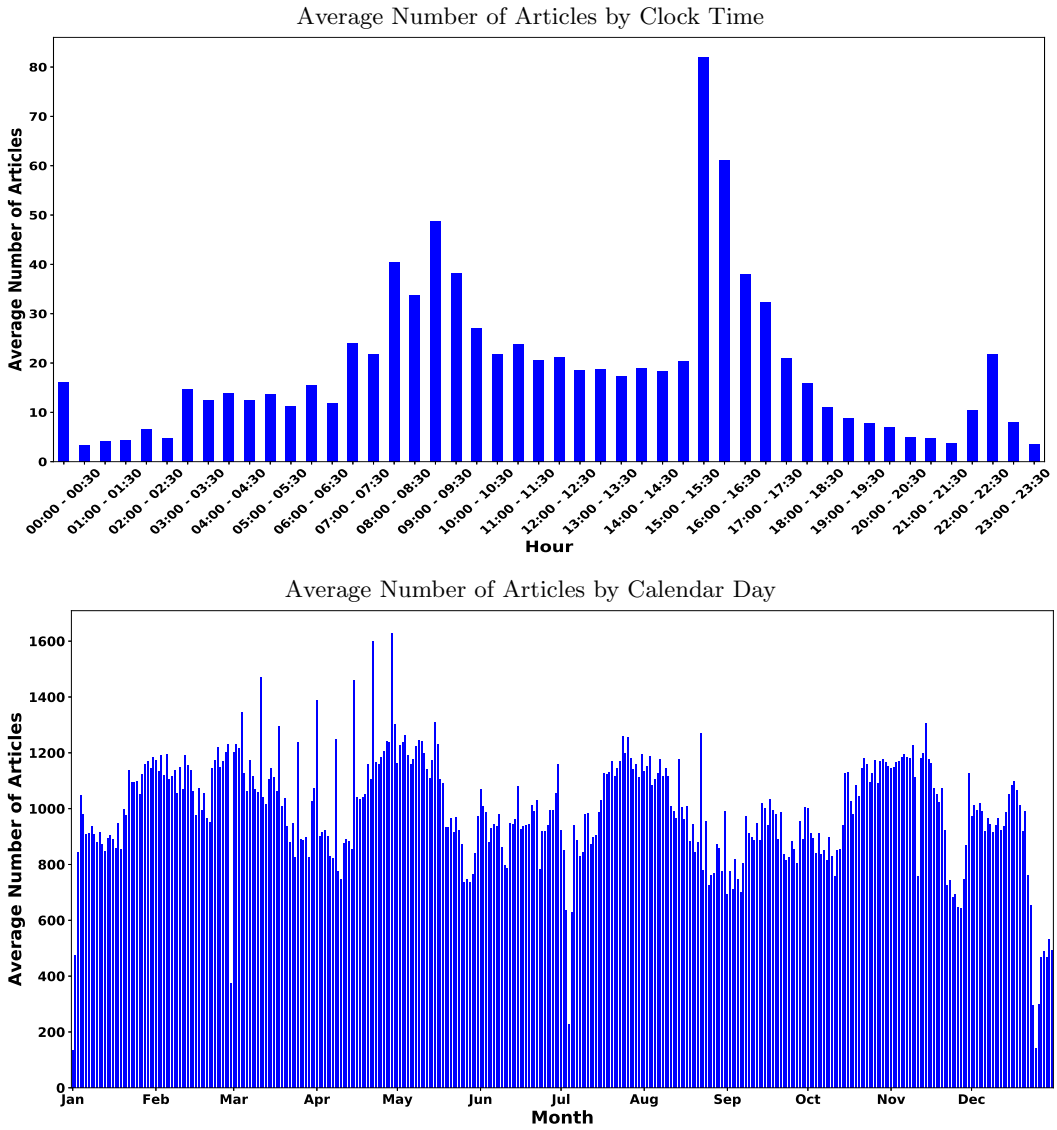
paths, at least over short horizons. While this alternative hypothesis is by now uncontroversial, it is hard to overstate its importance, as we have much to learn about the mechanisms through which information enters prices and the frictions that impede these mechanisms. Our prediction analysis adds new evidence to the empirical literature investigating the alternative hypothesis. In particular, we bring to bear information from a rich news text data set. Our methodological contribution is a new toolkit that makes it feasible to conduct a coherent statistical analysis of such complex and unstructured data. An ideal outcome of our analysis and future research using our method is to better understand how news influences investor belief formation and in turn enters prices.

4.1 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 body text 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 not only with contemporaneous returns but also with future returns.

Figure 2: Average Article Counts



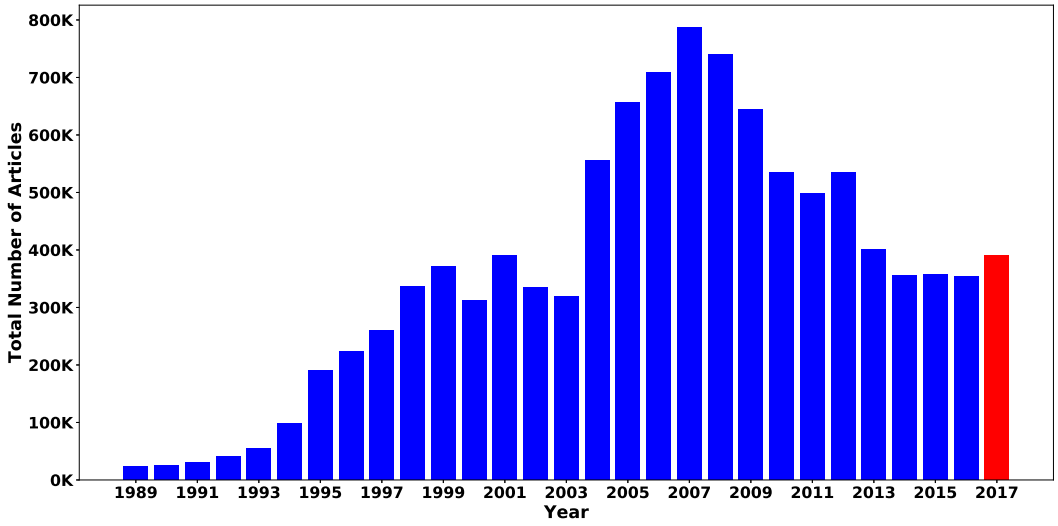
Note: The top figure plots the average numbers of articles per half an hour (24 hour EST time) from January 1, 1989 to July 31, 2017. The bottom figure plots the average numbers of articles per calendar day. Averages are taken over the full sample from January 1, 1987 to July 31, 2017.

Newswires are a highly visible information source for market participants, so presumably any delay in price response would be short-lived. Or, it could be the case that Newswires are a restatement of recently revealed information, in which case news is best aligned with prior returns.

Without better guidance on timing choice, we train the model by matching articles published on day t (more specifically, between 4pm of day $t-1$ and 4pm of day t) with the tagged firm's three-day return from $t-1$ to $t+1$ (more specifically, from market close on day $t-2$ to close on day $t+1$).[3]

[3]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.

Note that this timing is for sentiment training purposes only so as to achieve accurate parameter estimates. In order to devise a trading strategy, for example, it is critical to align sentiment estimates for an article *only* with future realized returns (we discuss this 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) 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 interval throughout the 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 articles until around 2007. Some news volume patterns reflect structural changes in news data sources and some reflect variation in the number of listed stocks. 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.

The first step is to remove proper nouns.[4] Next, we follow common steps from the natural language processing literature to clean and structure news articles.[5] 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.[6] The second step is stemming and 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.[7] 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."[8] 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.[9] 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.

4.2 Return Predictions

We train the model using rolling window estimation. The rolling window consists of a fifteen year interval, the first ten years of which are used for training and the last five years are 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 SESTM models corresponding to a grid of tuning parameters.[10] 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.

[4] We thank Timothy Loughran for this suggestion.

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

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

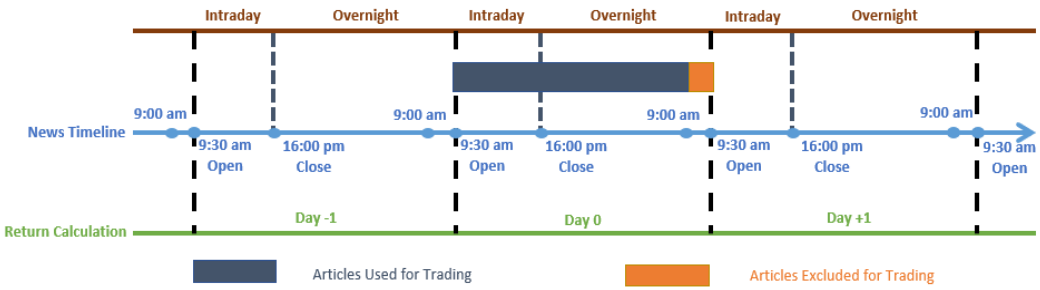
[7] The lemmatization procedure uses WordNet as a reference database: https://wordnet.princeton.edu/. The stemming procedure uses the package "porter2stemmer" on https://pypi.org/project/porter2stemmer/. Frequently, the stem of an English word is not itself an English word; for example, the stem of "accretive" 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.

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

[9] 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.

[10] 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).

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, . . . , day±10 portfolios.

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 reports 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 to 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 $\frac{1}{T}\sum_{t=1}^{T}\left(\sum_i \left|w_{i,t+1} - \frac{w_{i,t}(1+y_{i,t+1})}{\sum_j w_{j,t}(1+y_{j,t+1})}\right|\right)$, where $w_{i,t}$ is the weight of stock i in the portfolio at time t.

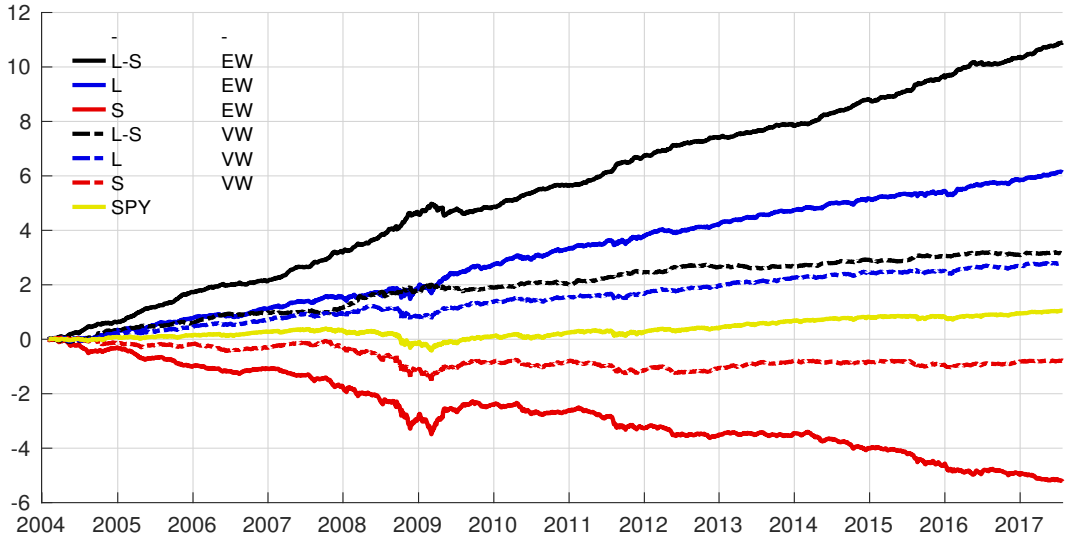
4.3 Daily Predictions

Figure 5 reports the cumulative one-day trading strategy returns (calculated from open-to-open) based on out-of-sample SESTM sentiment forecasts. We report the long (denoted "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.

To evaluate out-of-sample predictive performance in economic terms, we design a trading strategy

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



Note: This figure compares the out-of-sample cumulative log returns of portfolios sorted on 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).

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.[11]

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

[11] In the early part of the sample, there are a handful of days for which fewer than 50 firms have non-neutral scores, in which case we trade fewer than 100 stocks but otherwise maintain the zero-cost nature of the portfolio.

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 indicates 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 the fact that i) small stocks receive less investor attention and thus respond more slowly to news, ii) the underlying fundamentals of small stocks are more uncertain and opaque and thus it requires more effort to process news into actionable price assessments, or iii) 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.

Third, SESTM 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.

4.4 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 words' average sentiment tone $(O_+ - O_-)$ over all 14 training samples. Table A.2 in Appendix E provides additional detail on selected words, reporting the top 50 positive and negative sentiment words throughout our training 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 its average sentiment tone over all 14 training samples.

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

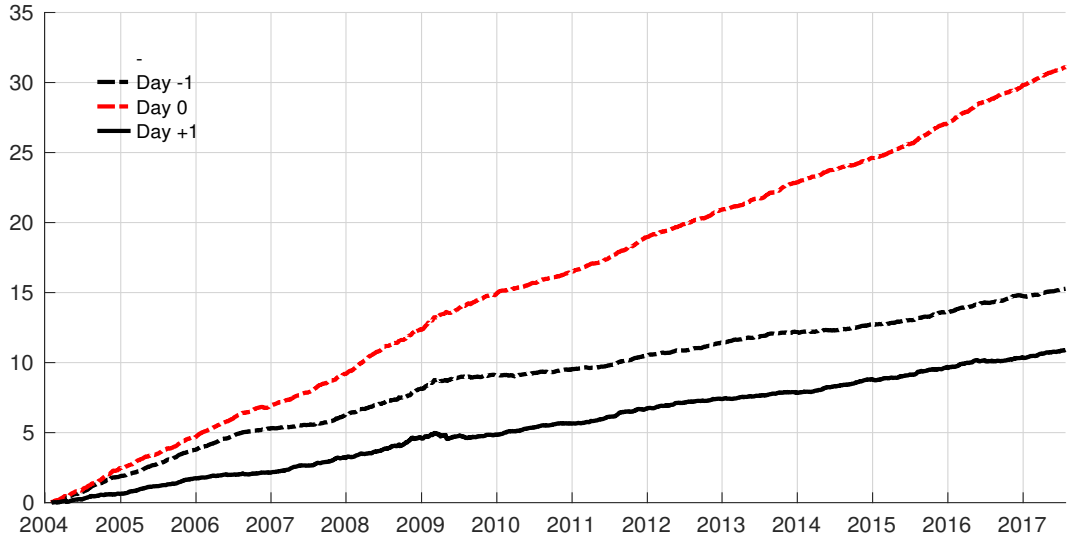
repurchase, surpass, upgrade, undervalue, surge, customary, jump, declare, rally, discretion, beat.

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 nine 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.

4.5 Speed of Information Assimilation

The analysis in Figure 5 and Table 2 focuses on relating news sentiment on day t to returns on day $t+1$. In the next two subsections, we investigate the timing of price responses to news sentiment with finer resolution.

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



Note: This figure compares the out-of-sample cumulative log returns of long-short portfolios sorted on 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.

4.5.1 Lead-lag Relationship Among News and Prices

In our training sample, we estimate SESTM 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 subsequent out-of-sample 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 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 generally 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, $\widehat{p}_i$, is based on a model estimated from an entirely distinct data set (that pre-dates the arrival of article i and returns $y_{i,t-1}$, $y_{i,t}$, and $y_{i,t+1}$). Table 3 reports summary statistics for these portfolios, including their annualized Sharpe ratios, average returns, alphas, and turnover. For this analysis, we specialize to equally weighted portfolios.

The Day −1 strategy (dashed black line) shows the association between news article sentiment

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

Formation	Sharpe Ratio	Turnover	Average Return	FF3 α	FF3 R^2	FF5 α	FF5 R^2	FF5+MOM α	FF5+MOM 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.

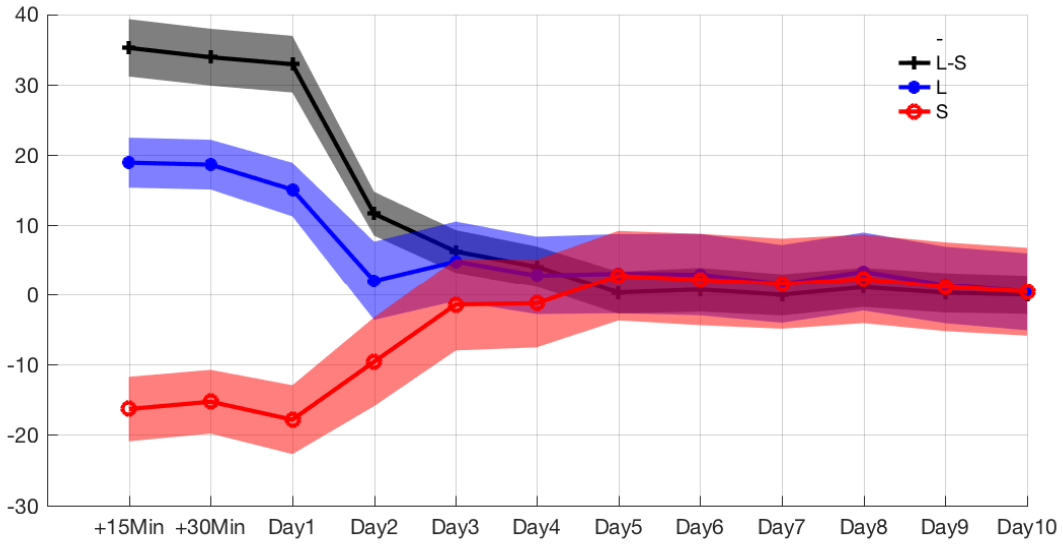
and the stock return one day prior to the news. This strategy thus quantifies the extent to which our sentiment score picks up on stale news. On average, prices move ahead of news in our sample, as indicated by the infeasible annualized Sharpe ratio of 5.88. Thus we see that much of the daily news flow echoes previously reported news or is a new report of information already known to market participants.

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 in news text. In particular, price responses are most concentrated on the same day that the news arrives, as reflected by the same-day infeasible annualized Sharpe ratio of 10.78.

The Day +1 strategy (solid black line) shows the association between news on day t and returns on the subsequent day. It 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 alternative hypothesis.

We next 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

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.

position either 15 or 30 minutes after the article's time stamp, and holds positions until the next day's open. We also study one-day open-to-open returns initiated anywhere from one to 10 days following the announcement.

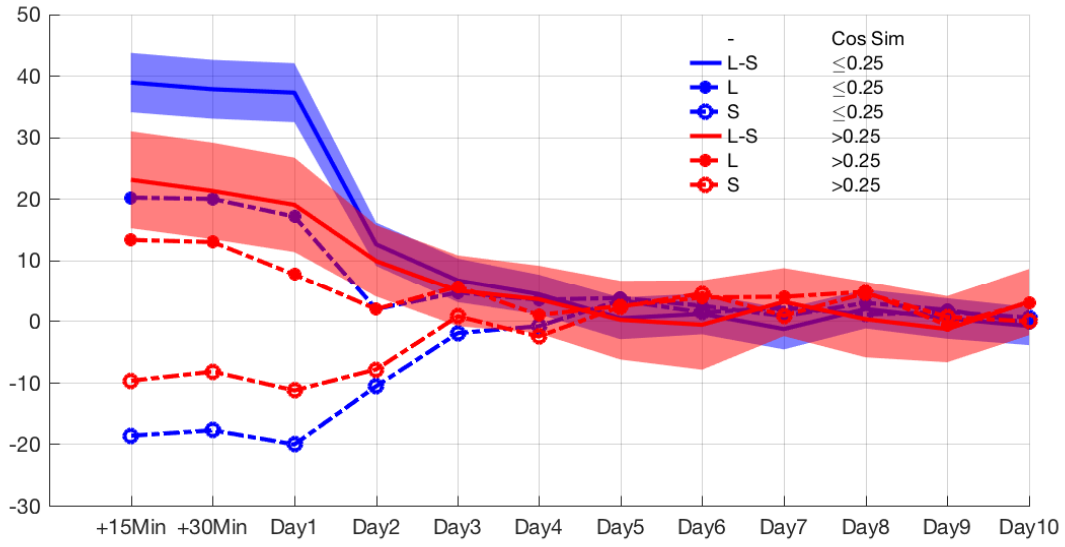
Figure 8 reports average returns in basis points per day with shaded 95% confidence intervals. It shows the long-short portfolio as well as the long and short legs separately. For the long-short strategy, sentiment information is essentially fully incorporated into prices by the start of Day +3. For 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.

4.5.2 Fresh News and Stale News

The evidence in Section 4.5.1 indicates that a substantial fraction of news is "old news" and already impounded in prices by the time an article is published. The assimilation analysis of Figure 8 thus pools together both fresh and stale news. In order to investigate the difference in price response to fresh versus stale news, we conduct separate analyses for articles grouped by the novelty of their content.

We construct a measure of article novelty as follows. For each article for firm i on day t, we

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.

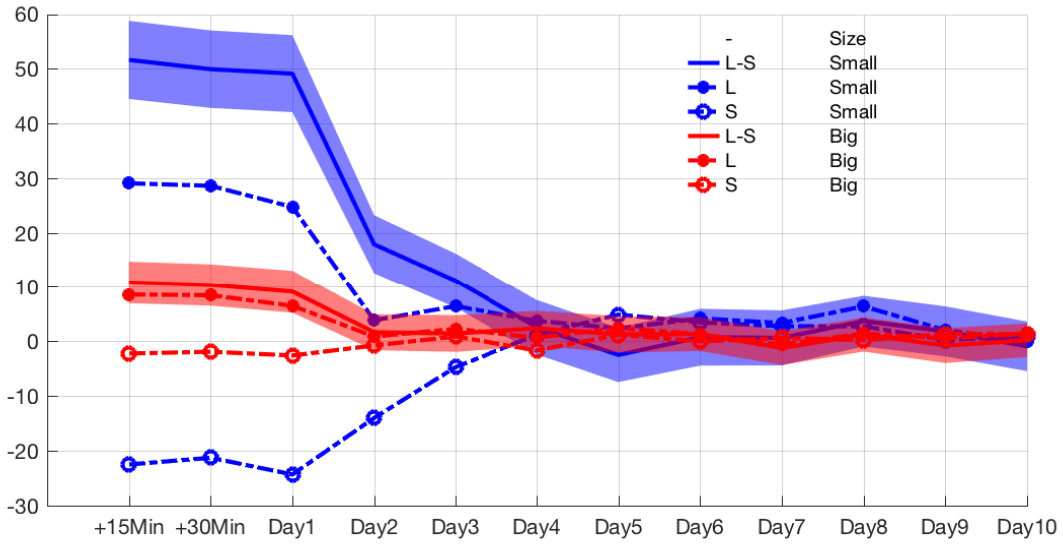
calculate its cosine similarity with all articles about firm i on the five trading days prior to t (denoted by the set $\chi_{i,t}$). Novelty of recent news is judged based on its most similar preceding article, thus we define article novelty as

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

Figure 9 splits out our news assimilation analysis by article novelty. We partition news into two groups. The "fresh" news group contains articles novelty score of 0.75 or more, while "stale" news has novelty below 0.75.[12] 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 23 basis point response to stale news (solid red line). Furthermore, it takes four days for fresh news to be fully incorporated in prices (i.e., the 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.

[12]The average article novelty in our sample is approximately 0.75. The conclusions from Figure 9 are generally insensitive to 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.

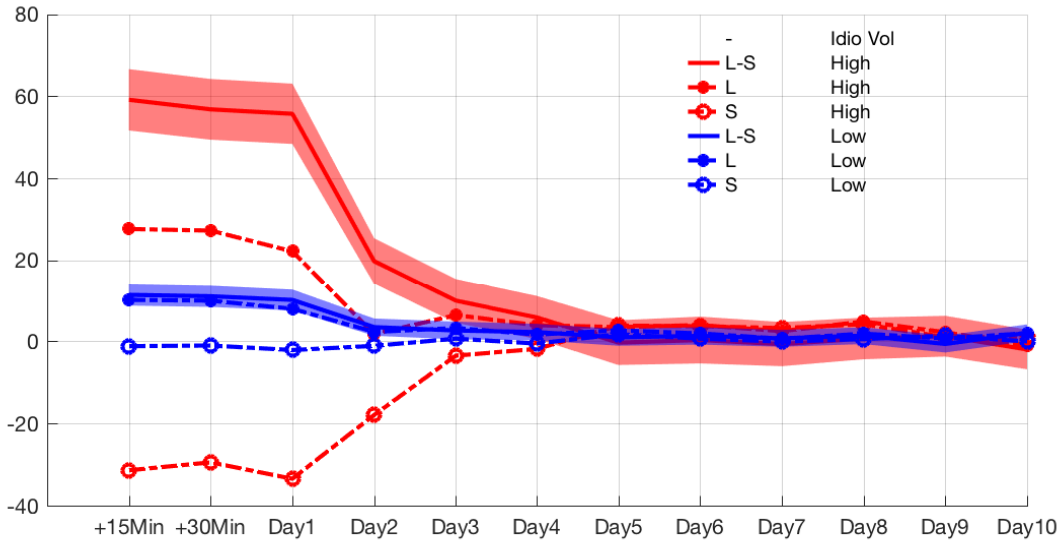
4.6 Stock Heterogeneity Analysis: Size and Volatility

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 a 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 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. 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

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.

price assimilation is ambiguous.

Figure 11 shows the comparative price response of high versus low volatility firms.[13] The price response to SESTM sentiment in the first 15 minutes following news arrival is 12 basis points for low volatility firms, but 59 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 reflected in the price of a high volatility stock.

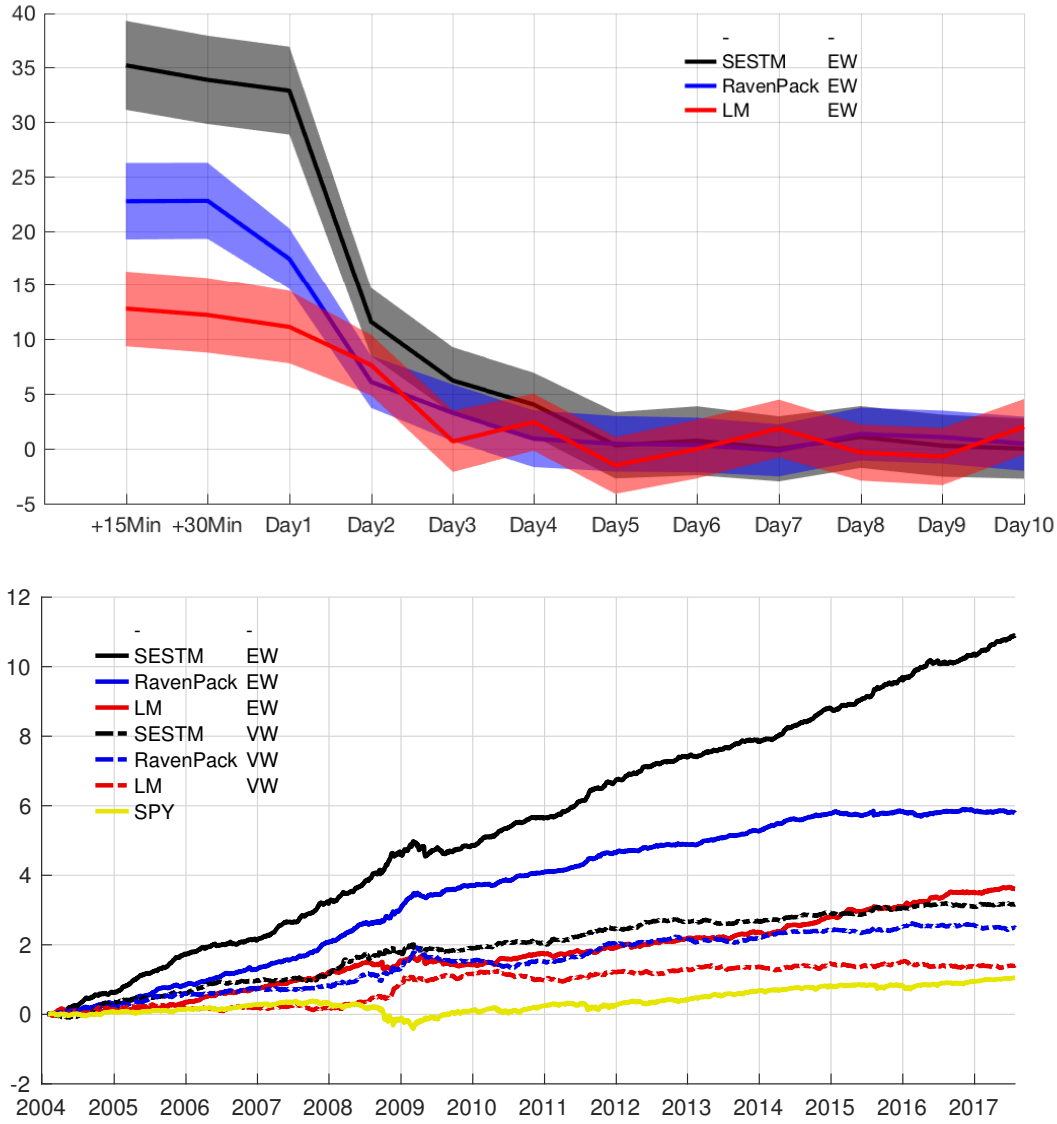
4.7 Comparison Versus Dictionary Methods and RavenPack

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

The first alternative for comparison is dictionary-based sentiment scoring. We construct the LM sentiment score of an article 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 in their negative dictionary. As with SESTM, we average scores from multiple articles for the same firm in the same day. This produces a stock-day signal, $\widehat{p}_i^{LM}$, which we use to construct trading strategies in the same manner as the SESTM-based signal, $\widehat{p}_i^{SESTM}$, in

[13] 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 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 12: SESTM Versus LM and RavenPack



Note: For top panel notes, see Figure 8. In addition to SESTM, 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 SESTM, 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).

preceding analyses.

The second alternative for comparison are news sentiment scores from RavenPack News Analytics 4 (RPNA4). As stated on its website,[14]

> RavenPack is the leading big data analytics provider for financial services. Financial professionals

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

Table 4: SESTM Versus LM and RavenPack

	Sharpe		Average	FF6+SESTM			FF6+LM			FF6+RP		
EW/VW	Ratio	Turnover	Return	α	$t(\alpha)$	R^2	α	$t(\alpha)$	R^2	α	$t(\alpha)$	R^2
					SESTM							
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 day.

> 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. 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 SESTM (*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 D. Following the same procedure used for $\widehat{p}_i^{SESTM}$ and $\widehat{p}_i^{LM}$, we construct RavenPack daily stock-level sentiment scores ($\widehat{p}_i^{RP}$) by averaging all reported article sentiment scores pertaining to a given firm in a given day.[15]

We build trading strategies using each of the three sentiment scores, $\widehat{p}_i^{SESTM}$, $\widehat{p}_i^{LM}$, and $\widehat{p}_i^{RP}$. 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 12 assesses the extent and timing of price responses for 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 SESTM is significantly more effective than alternatives in identifying price-relevant content of news articles. Beginning fifteen minutes after news arrival,

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

the one-day long-short return based on SESTM 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 (the effect of RavenPack is borderline insignificant at three days). The SESTM signal, on the other hand, requires four days to be fully incorporated in prices. This suggests that SESTM 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 12 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 SESTM 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 SESTM 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 12, as well as their daily turnover. The SESTM 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, SESTM earns an annualized Sharpe ratio of 4.3, versus 3.2 and 1.7 for RavenPack and LM, respectively. The outperformance of SESTM is also evident when comparing value-weighted Sharpe ratios. In this case, SESTM achieves a Sharpe ratio of 1.3 versus 1.1 for RavenPack and 0.7 for LM.

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 and 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 SESTM regressed on FF6 and the RavenPack strategy. The SESTM α's are in each case almost as large as its raw return. At most, 15% of the SESTM 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 for the alternative strategies after controlling for SESTM, indicating not only that they achieve significant positive returns, but also that a component of those excess returns are uncorrelated with SESTM and FF6. In short, SESTM, RavenPack, and LM capture different varieties of information content

in news articles, which suggests potential mean-variance gains from combining the three strategies. Indeed, a portfolio that places one-third weight on each of the equal-weight sentiment strategies earns an annualized out-of-sample Sharpe ratio of 4.9, significantly exceeding the 4.3 Sharpe ratio of SESTM on its own.

4.8 Transaction Costs

Our trading strategy performance analysis thus far ignores transaction costs. This is because the portfolios above 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 SESTM, LM, and RavenPack is apples-to-apples in the sense that all three strategies face the same trading cost environment.

That said, evaluating the usefulness of news article sentiment for practical portfolio choice is a separate question and is interesting in its own right. However, the practical viability of our sentiment strategies is difficult to ascertain from preceding tables due to their large turnover. In this section, to better understand the relevance of SESTM's predictability gains for practical asset management, we investigate the performance of sentiment-based trading strategies while taking into account trading costs.

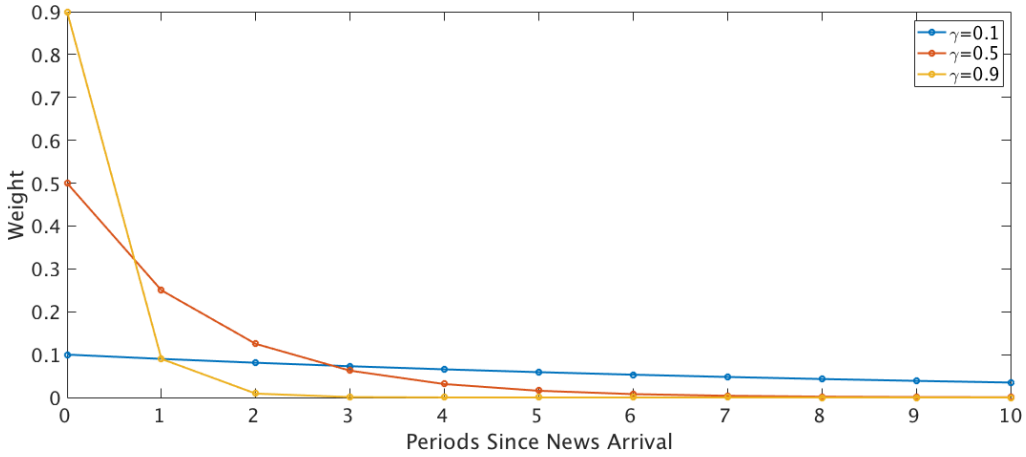
To approximate the net performance of a strategy, we assume that each portfolio incurs a daily transaction cost of 10bps. The choice of 10bps approximates the average trading cost experienced by large asset managers, as reported in Frazzini et al. (2018).

We propose a novel trading strategy that directly reduces portfolio turnover and hence trading costs. Specifically, we design a strategy that i) turns over (at most) a fixed proportion of the existing portfolio every period and ii) assigns weights to stocks that decay exponentially with the time since the stock was in the news. These augmentations effectively extend the stock holding period from one day to multiple days. We refer to this approach as an exponentially-weighted calendar time (EWCT) portfolio.

On the first day of trading, we form an equal-weighted portfolio that is long the top N stocks in terms of news sentiment that day and short N stocks with the most negative news sentiment. A single parameter (γ) determines the severity of the turnover constraint. Each subsequent day t, we liquidate a fixed proportion γ of all existing positions, and reallocate that γ proportion to an equal-weighted long-short portfolio based on day t news. For a stock i experiencing large positive sentiment news on day t, its weight changes according to $w_{i,t} = \frac{\gamma}{N} + (1-\gamma)w_{i,t-1}$. For a stock i in the long-side of the portfolio at day $t-1$ but with no news on date t, its portfolio weight decays to $w_{i,t} = (1-\gamma)w_{i,t-1}$. The analogous weight transitions apply to the short leg of the strategy.

To see this more clearly, consider an example with three stocks, A, B, and C, in a broader cross section of stocks. Suppose at time t that A has a weight of zero ($w_{A,t} = 0$) while B and C had their first and only positive news five and ten days prior, respectively (that is, $w_{B,t} = (1-\gamma)^4\gamma/N$ and $w_{C,t} = (1-\gamma)^9\gamma/N$). Now suppose that, at time $t+1$, positive news articles about stocks A and C

Figure 13: EWCT Weight Decay



Note: Illustration of portfolio weight decay in the turnover-constrained EWCT trading strategy.

propel them into the long side of the sentiment strategy, and neither A, B, nor C experiences news coverage thereafter. The weight progression of A, B, and C is the following:

	t	$t+1$	$t+2$	$t+3$	...
w_A	0	$\frac{\gamma}{N}$	$\frac{\gamma}{N}(1-\gamma)$	$\frac{\gamma}{N}(1-\gamma)^2$	...
w_B	$\frac{\gamma}{N}(1-\gamma)^4$	$\frac{\gamma}{N}(1-\gamma)^5$	$\frac{\gamma}{N}(1-\gamma)^6$	$\frac{\gamma}{N}(1-\gamma)^7$	...
w_C	$\frac{\gamma}{N}(1-\gamma)^9$	$\frac{\gamma}{N}\left(1+(1-\gamma)^{10}\right)$	$\frac{\gamma}{N}(1-\gamma)\left(1+(1-\gamma)^{10}\right)$	$\frac{\gamma}{N}(1-\gamma)^2\left(1+(1-\gamma)^{10}\right)$	...

The portfolio weights for A and C spike upon news arrival and gradually revert to zero. The turnover parameter simultaneously governs both the size of the weight spike at news arrival (the amount of portfolio reallocation) as well as the exponential decay rate for existing weights. This is illustrated in Figure 13. For high values of γ, new information is immediately assigned a large weight in the portfolio and old information is quickly discarded, generating large portfolio turnover. In contrast, low values of γ reduce turnover both by limiting the amount of wealth reallocated to the most recent news and by holding onto past positions for longer, which in turn increases the effective holding period of the strategy. Finally, note that the EWCT strategy guarantees daily turnover is never larger than γ. When a stock is already in a portfolio and a new article arrives with the same sign as recent past news (as in the example of stock C) the actual turnover will be less than γ.

Table 5 reports the performance of EWCT portfolios as we vary turnover limits from mild ($\gamma = 0.9$) to heavily restricted ($\gamma = 0.1$). Moving down the rows we see that a more severe turnover restriction drags down the gross Sharpe ratio of the trading strategy, indicating a loss in predictive information due to signal smoothing. This drag is offset by a reduction in trading costs. As a result, the net Sharpe ratio peaks at 2.3 when $\gamma = 0.5$. That is, with a moderate amount of turnover control (and concomitant signal smoothing), the gain from reducing transaction costs outweighs the loss in predictive power. In sum, Table 5 demonstrates the attractive risk-return tradeoff to investing based on news sentiment even after accounting for transactions costs.

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

		Gross		Net	
γ	Turnover	Return	Sharpe Ratio	Return	Sharpe Ratio
0.1	0.08	5.18	1.77	3.58	1.17
0.2	0.17	9.74	2.93	6.31	1.84
0.3	0.27	13.71	3.61	8.37	2.16
0.4	0.36	17.24	4.03	9.98	2.28
0.5	0.46	20.43	4.26	11.23	2.30
0.6	0.56	23.32	4.38	12.17	2.25
0.7	0.66	25.97	4.43	12.88	2.15
0.8	0.75	28.43	4.42	13.39	2.04
0.9	0.85	30.74	4.37	13.74	1.92

Note: The table reports the performance of equally-weighted long-short EWCT portfolios based on SESTM scores. The EWCT parameter is γ. Average returns are reported in basis points per day and Sharpe ratios are annualized. Portfolio average daily turnover is calculated as $\frac{1}{T}\sum_{t=1}^{T}\left(\sum_i \left|w_{i,t+1} - \frac{w_{i,t}(1+y_{i,t+1})}{1+\sum_j w_{j,t}y_{j,t+1}}\right|\right)$.

5 Statistical Theory

We study the statistical properties of SESTM in an asymptotic framework. We recall that n is the number of training articles, m is the size of full vocabulary, $|S|$ is the size of sentiment dictionary, and $s_1, s_2, \ldots, s_n$ are the counts of all sentiment-charged words in individual training articles. We let n be the driving asymptotic parameter and allow $(m, |S|, s_1, \ldots, s_n)$ to depend on n. The probabilistic model of Section 2, summarized in equations (1) and (3), describes the distribution of training data given sentiment scores $p_1, p_2, \ldots, p_n$. We treat these scores as (non-random) parameters.

5.1 Model Assumptions

First, we formally characterize the difference between sentiment charged words and sentiment neutral ones. For sentiment charged words, their corresponding entries in $O_\pm$ should be different. Otherwise, these words would not represent any sentiment and should be left in set N. At the same time, the sentiment neutral words are analogous to useless predictors in a linear model, for which the regression coefficients are zero. We introduce a similar quantity here:

$$\delta_j = \mathbb{E}\Big[n^{-1}\sum_{i=1}^{n}(\mathrm{sgn}(y_i) - \overline{\mathrm{sgn}}_y)\, d_{j,i}\Big], \qquad \text{for } 1 \le j \le m,$$

where $\overline{\mathrm{sgn}}_y = n^{-1}\sum_{i=1}^{n}\mathrm{sgn}(y_i)$. This quantity measures the strength of association between the count of a word and the sign of tagged return. As we will show in the proof of Theorem 1, $O_{+,j} - O_{-,j} \neq 0$ implies $\delta_j \neq 0$ for any $j \in S$. For $j \in N$, there is no $O_{\pm,j}$ defined. We directly impose the assumption $\delta_j = 0$ to differentiate a word in N from words in S.

Assumption 1 (Word Classification)**.** *For any $j \in S$, $O_{+,j} - O_{-,j} \neq 0$; for any $j \in N$, $\delta_j = 0$.*

Next, we discuss model identification. As we mentioned earlier, part of this model, (3), is a

two-topic topic model. Topic models are not identifiable without additional restrictions.[16] To see this, it follows by direct calculations that for all $1 \le i \le n$, we have

$$p_i O_+ + (1-p_i)O_- = \widetilde{p}_i \widetilde{O}_+ + (1-\widetilde{p}_i)\widetilde{O}_-,$$

where $\widetilde{O}_+ = (1-\alpha)O_+ + \alpha O_-$, $\widetilde{O}_- = (1+\alpha)O_- - \alpha O_+$, and $\widetilde{p}_i = p_i + \alpha$, for any α such that all $\widetilde{p}_i$s are within $[0,1]$, all entries of $\widetilde{O}_\pm$ are non-negative, and $\widetilde{O}_{+,j} - \widetilde{O}_{-,j} \neq 0$ for any j.

In light of this, we impose the following restriction on p_is to remove this extra degree of freedom, which is amount to a requirement that the average tone for all news is neutral:

Assumption 2 (Score Normalization)**.** $n^{-1}\sum_{i=1}^n p_i = 1/2$.[17]

The methodology in Section 2 only needs the marginal distributions of y_i and $d_{[S],i}$, so we leave the full distribution of data unmodeled. To provide theoretical justification, we need mild assumptions on the joint distribution of $\{(y_i, d_{[S],i}, d_{[N],i})\}_{1\le i\le n}$.

We first discuss the assumption on the dependence among $\{y_1, y_2, \ldots, y_n\}$. The collection of returns in the training sample contain many stocks across different time periods. Because of the potential cross-sectional dependence and time dependence, it is inappropriate to use conventional assumptions such as strong mixing. We adopt the notion of *dependency graph* (Janson, 2004) to describe the dependence among y_is. Given the joint distribution of $\{y_i\}_{1\le i\le n}$, an undirected graph Γ with nodes $\{1, 2, \ldots, n\}$ is called a dependency graph if, for any $i \in \{1, 2, \ldots, n\}$ and $V \subset \{1, 2, \ldots, n\}\backslash\{i\}$, there is no edge between i and nodes in V implies that y_i is independent of $\{y_j\}_{j\in V}$. We shall assume there exists a non-trivial dependency graph whose maximum degree is properly small. In the special case where y_is follow a multivariate normal distribution, the support of the covariance matrix naturally defines a valid dependency graph, so the above assumption translates to the row-wise sparsity assumption on the covariance matrix.

We next consider the conditional distribution of $\{(d_{[S],i}, d_{[N],i})\}_{1\le i\le n}$ given $\{y_i\}_{1\le i\le n}$. Let s_i and n_i denote the total count of words from S and N in article i, respectively. We adopt the conventional bag-of-words model by assuming the s_i words (n_i words) are *i.i.d.* drawn from the word list S (word list N) according to some distribution $Q_i \in \mathbb{R}^{|S|}$ ($\Omega_i \in \mathbb{R}^{|N|}$). Equivalently,

$$d_{[S],i}\big|\{y_1, \ldots, y_n\} \sim \text{Multinomial}(s_i, Q_i), \qquad d_{[N],i}\big|\{y_1, \ldots, y_n\} \sim \text{Multinomial}(n_i, \Omega_i).$$

The dependence on y_is is reflected in the probability vectors $Q_i = Q_i(y_1, \ldots, y_n)$ and $\Omega_i = \Omega_i(y_1, \ldots, y_n)$. We assume $Q_i(y_1, \ldots, y_n) = p_i O_+ + (1-p_i)O_-$, which is compatible with the marginal distribution of $d_{[S],i}$ in (3). Regarding $\Omega_i(y_1, \ldots, y_n)$, we make a mild assumption that $\Omega_i(y_1, \ldots, y_n) = \Omega_i(y_i)$. To summarize:

Assumption 3 (Joint Distribution)**.** *The following statements hold:*

[16] Such restrictions in the literature include a parametric Dirichlet model for topic weights (Blei et al., 2003) and the existence of anchor words (Arora et al., 2012).

[17] We note that this assumption has no conflict with unequal numbers of positive and negative returns in the training data. Recall that g is as in (1). If $g(1/2) > 1/2$, then the average tone is neutral yet there are more than half of training articles with positive returns.

(a) There exists a valid dependency graph for $\{y_i\}_{1\le i\le n}$ *whose maximum degree* $K_n = o(n/\log^2(m))$.

(b) Conditioning on $\{y_i\}_{1\le i\le n}$*, the random vectors* $\{d_{[S],1},\ldots,d_{[S],n},d_{[S],1},\ldots,d_{[N],n}\}$ *are independent,* $d_{[S],i} \sim \text{Multinomial}\big(s_i,\, p_iO_+ + (1-p_i)O_-\big)$*, and* $d_{[N],i} \sim \text{Multinomial}\big(n_i, \Omega_i(y_i)\big)$.

Finally, we need some regularity conditions. Let s_i and n_i be the same as in Assumption 3. Denote by $s_{\max}$, $s_{\min}$, and $\bar{s}$ the maximum, minimum, and average of $\{s_i\}_{1\le i\le n}$, respectively, and $n_{\max}$, $n_{\min}$, and $\bar{n}$ are defined accordingly. Let $\Omega_{j,i}(y_i)$ denote the j^{th} entry of the vector $\Omega_i(y_i)$, for $j \in N$.

Assumption 4 (Regularity Conditions)**.** *There exist constants* $c_0 > 0$, $c_1 > 0$, *and* $C > 0$ *such that the following statements hold:*

(a) $s_{\max} \le Cs_{\min}$*, and* $n_{\max} \le Cn_{\min}$.

(b) For each $j \in N$*, with probability approaching* 1, $|\Omega_{j,i}(y_i)| \le Cq_j$*, where* $q_j = n^{-1}\sum_{i=1}^n \mathbb{E}[\Omega_{j,i}(y_i)]$.

(c) $\min\big\{n\bar{s}\min_{j\in S}(O_{+,j}+O_{-,j}),\ n\bar{n}\min_{j\in N} q_j\big\}/\log(m) \to \infty$.

(d) $\sum_{i=1}^n s_i[p_iO_{+,j} + (1-p_i)O_{-,j}] \ge c_0\sum_{i=1}^n s_i(O_{+,j}+O_{-,j})$*, for any* $j \in S$.

(e) $|\sum_{i=1}^n (g(p_i)-\bar{g})(s_i-\bar{s})| \le C\bar{s}\sqrt{nK_n\log(m)}$*, where* $\bar{g} = n^{-1}\sum_{i=1}^n g(p_i)$ *and* K_n *is the same as in Assumption 3.*

Condition (a) requires s_i's to be of the same order; similar for n_i's. This condition ensures the theoretical result is not driven by some articles with substantially more or less words than others. Condition (b) prevents the frequency of a sentiment-neutral word j in any particular article to be much larger than its average frequency in all articles. In Condition (c), the expected counts of a word in all training articles are of the order $n\bar{s}(O_{+,j}+O_{-,j})$ for a sentiment-charged word, and $n\bar{n}q_j$ for a sentiment neutral word, which dominate $\log(m)$, because n is usually very large in real data. Condition (d) and Condition (e) are high-level technical conditions. The former is satisfied when all p_i's are bounded away from 0 and 1, whereas the latter holds with probability approaching one if we assume s_i's are independently and identically distributed.

5.2 Consistency of Screening

The screening step estimates the sentiment dictionary by $\widehat{S}$. We establish the consistency of screening by showing that $\hat{S} = S$ with an overwhelming probability. This property is analogous to the property of model selection consistency for linear models (see, e.g. Fan and Lv, 2008).

We define a quantity to capture the *sensitivity of stock returns to article sentiment*:

$$\theta \equiv \frac{\sum_{i=1}^n s_i\big(p_i - 1/2\big)\big[g(p_i)-\bar{g}\big]}{\sum_{i=1}^n s_i}, \tag{12}$$

where $g(\cdot)$ is the monotone increasing function in Model (1) and $\bar{g} = n^{-1}\sum_{i=1}^n g(p_i)$. θ captures the steepness of g and the extremeness of training articles' tones.

Lemma 1 (Screening Statistics). *Consider Models (1)-(3), where Assumptions 1-4 are satisfied. Let ϵ_n be a sequence such that $\epsilon_n = o(1)$ and $\log(\epsilon_n^{-1}) = o(n \min\{K_n^{-1},\ \bar{s}\min_{j\in S}(O_{+,j}+O_{-,j}),\ \bar{n}\min_{j\in N} q_j\})$, where $(K_n, \bar{s}, \bar{n}, q_j)$ are the same as in Assumption 4. As $n \to \infty$, with probability $1-\epsilon_n$, simultaneously for all $1 \le j \le m$, we have*

$$|f_j - \widehat{\pi}| \begin{cases} \ge 2\theta \frac{|O_{+,j}-O_{-,j}|}{O_{+,j}+O_{-,j}} - \frac{C\sqrt{\log(m/\epsilon_n)}}{\sqrt{n\min\{K_n^{-1},\ \bar{s}(O_{+,j}+O_{-,j})\}}}, & \text{when } j \in S, \\ \le \frac{C\sqrt{\log(m/\epsilon_n)}}{\sqrt{n\min\{K_n^{-1},\ \bar{n}q_j\}}}, & \text{when } j \in N. \end{cases}$$

In the screening step, $\widehat{S}$ is obtained by thresholding $|f_j - \widehat{\pi}|$. Theorem 1 justifies that $|f_j - \widehat{\pi}|$ is relatively large for sentiment-charged words and relatively small for sentiment-neutral words. For any choice of $\epsilon_n = o(1)$, the statement of this lemma holds with probability approaching 1. In what follows, we assume $m \to \infty$ as $n \to \infty$ and fix $\epsilon_n = m^{-1}$; item (c) in Assumption 4 ensures that $\epsilon_n = m^{-1}$ satisfies the requirement of Lemma 1. This choice of ϵ_n greatly simplifies the expressions. When m is finite, we only need to replace $\log(m)$ in the statements below by $\log(m) + \log(1/\epsilon_n)$ for a proper sequence $\epsilon_n = o(1)$.

The next theorem establishes the consistency of screening:

Theorem 1 (Consistency of Screening). *Consider Models (1)-(3), where Assumptions 1-4 are satisfied. Suppose the following condition holds:*

$$n\theta^2 \min_{j\in S} \frac{(O_{+,j}-O_{-,j})^2}{(O_{+,j}+O_{-,j})^2} \ge \frac{\log^2(m)}{\min\{K_n^{-1},\ \bar{s}\min_{j\in S}(O_{+,j}+O_{-,j}),\ \bar{n}\min_{j\in N} q_j\}}. \tag{13}$$

Now set $\kappa = \log(m)$ and $\alpha_\pm = \frac{\sqrt{\log(m)\log(\log(m))}}{\sqrt{n\min\{K_n^{-1}, \bar{s}\min_{j\in S}(O_{+,j}+O_{-,j}), \bar{n}\min_{j\in N} q_j\}}}$ in (6). As $n, m \to \infty$, we have $\mathbb{P}(\widehat{S} = S) = 1 - o(1)$.

To obtain consistency for screening, the desired number of training articles, n, is determined by three factors: 1) sensitivity of stock return to article sentiment, θ, defined in (12); 2) $\min_{j\in S} \frac{|O_{+,j}-O_{-,j}|}{O_{+,j}+O_{-,j}}$, the minimum among frequency-adjusted sentiments of all words in S; 3) $\min\{K_n^{-1}, \bar{s}\min_{j\in S}(O_{+,j}+O_{-,j}), \bar{n}\min_{j\in N} q_j\}$, in which the last two terms inside 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 K_n^{-1}, the third factor is equal to K_n^{-1}. For short articles, the per-article count of a word may approach zero faster than K_n^{-1}, in which case more training articles are needed, as shown from this theorem.

5.3 Accuracy of Learning Sentiment Topics

We quantify the estimation errors on two sentiment vectors: $O_{\pm}$. Since they can be re-parametrized by the vector of frequency, F, and the vector of tone, T, via (4), we consider the estimators

$$\widehat{F} = \frac{1}{2}(\widehat{O}_+ + \widehat{O}_-), \qquad \widehat{T} = \frac{1}{2}(\widehat{O}_+ - \widehat{O}_-), \tag{14}$$

and quantify the estimation errors for F and T. The error bounds we present below can certainly translate to those for $O_{\pm}$.

The estimation step obtains $\widehat{O}_+$ and $\widehat{O}_-$ by plugging in a "crude" estimate $\widehat{p}_i$ for p_i; see (9). The estimation error in this step naturally propagates. Importantly, we quantify the quality of the approximation of $\{p_i\}_{i=1}^n$ by $\{\widehat{p}_i\}_{i=1}^n$:

$$\widehat{\rho} = \frac{12}{n}\sum_{i=1}^{n}\Big(p_i - \frac{1}{2}\Big)\Big(\widehat{p}_i - \frac{1}{2}\Big). \tag{15}$$

Here, $\widehat{\rho}$ can be viewed as a particular form of "correlation" between the sentiment score and the normalized rank of return. In view of Model (1), it is reasonable to assume that $\widehat{\rho} > 0$.

The next theorem presents our main result about the estimation step:

Theorem 2 (Estimation Errors for Sentiment Vectors)**.** *Consider Models (1)-(3), where Assumptions 1-4 hold and (13) is satisfied. Suppose $0 < \widehat{\rho} \le 1$. As $n, m \to \infty$, with probability approaching 1,*

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

A few points are worth mentioning. Theorem 2 states that our method in fact estimates $(F, \widehat{\rho}T)$ rather than (F, T). This is the price we pay by not imposing any particular model on the distribution of returns. Suppose, instead of Model (1), we specify a parametric distribution of y_i given p_i, then we could construct more accurate estimates of p_i such that $n^{-1}\sum_{i=1}^n(\widehat{p}_i - p_i)^2 \to 0$, which in turn results in the consistency for the estimation of T. This is confirmed by the next corollary:[18]

Corollary 1. *Suppose conditions of Theorem 2 hold. Let $\widehat{p}_i$ be an arbitrary estimate of p_i, including but not limited to the $\widehat{p}_i$ in (9). We obtain $\widehat{O}_{\pm}$ similarly as in (10), except that $\widehat{W}$ is constructed from the new $\widehat{p}_i$'s. Suppose $\lambda_{\min}(n^{-1}\widehat{W}\widehat{W}') \ge a_0$, for a constant $a_0 > 0$. As $n, m \to \infty$, with probability approaching 1,*

$$\max\big\{\|\widehat{F} - F\|_1,\, \|\widehat{T} - T\|_1\big\} \le C\sqrt{\frac{|S|\log(m)}{n\bar{s}}} + C\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{p}_i - p_i)^2\right]^{1/2}.$$

Consequently, if $n^{-1}\sum_{i=1}^n(\widehat{p}_i - p_i)^2$ converges to 0 in probability, then $\|\widehat{F} - F\|_1$ and $\|\widehat{T} - T\|_1$ also diminish with probability approaching 1.

[18] This corollary applies to any estimate $\widehat{p}_i$. For the particular $\widehat{p}_i$ in (9), due to its special structure, the rate of convergence for $\|\widehat{F} - F\|_1$ is faster without the additional term on the right-hand side; see Theorem 2.

We prefer not to make stronger assumptions as in Corollary 1, because our empirical analysis does not require the consistency for T. In fact, despite its potential inconsistency, $\widehat{T}$ preserves the *sign, order, and relative strength of tone for words*. Mathematically speaking, the sign of T_j indicates whether word j is positive or negative. For two positive words j and k, the ratio T_j/T_k describes the strength of tone of word j relative to word k. Since $\widehat{\rho} > 0$, we have

$$\operatorname{sgn}(T_j) = \operatorname{sgn}((\widehat{\rho}T)_j), \qquad \text{and} \qquad (\widehat{\rho}T)_j \,/\, (\widehat{\rho}T)_k = T_j \,/\, T_k, \qquad \text{for all } (j,k).$$

This property plays a key role in the next step of scoring new articles (see Section 5.4).

Last but not least, we compare our results with those from an unsupervised topic modeling approach by Ke and Wang (2017). They propose a singular value decomposition approach to recover topic vectors $\widehat{O}_{\pm}$ using $D_{[\hat{S}],\cdot}$ alone without any supervision from returns. While this unsupervised approach can achieve consistency in terms of $\|\widehat{T} - T\|_1$, this is inconsequential in our context. As we have seen above and will learn from Section 5.4, an accurate estimate of ρT is good enough for the purpose of scoring new articles. On the other hand, the unsupervised approach suffers a much slower rate of convergence. According to Table 5 of Ke and Wang (2017), the best attainable error rate for any unsupervised estimator is

$$\sqrt{\frac{|S|\log(m)}{n\bar{s}}}\left(1 + \frac{|S|}{\bar{s}}\right), \qquad \text{up to some logarithmic factor.}$$

Recall that $|S|$ is the size of sentiment dictionary and $\bar{s}$ is the average per-article count of sentiment-charged words. In most real applications, $|S| \gg \bar{s}$. For example, in our empirical study of the *Dow Jones Newswire* database, $|S|$ is roughly a few hundreds whereas $\bar{s}$ is typically below 10. Therefore, the bias-variance trade-off clearly favors the supervised approach we propose here.

5.4 Accuracy of Scoring New Articles

Given the word count vector $d \in \mathbb{R}_+^m$ of a new article, we use (11) to predict its true sentiment score p. Theorem 3 quantifies the prediction accuracy of a single article's sentiment, whereas Theorem 4 investigates the accuracy of sentiment "ranks" for multiple articles.

Recall that $\widehat{\rho}$, as defined in (15), is related to the bias of $\widehat{T}$ in Theorem 2. As the estimation error propagates, a similar bias occurs to the estimator of p. We thereby define the *rescaled sentiment* score p^* as:

$$p^* = \begin{cases} 0, & \text{if } p < (1-\widehat{\rho})/2, \\ 1, & \text{if } p > (1+\widehat{\rho})/2, \\ 1/2 + (p - 1/2)/\widehat{\rho}, & \text{otherwise.} \end{cases} \tag{16}$$

Apparently, p^* is a monotonic increasing transformation of p such that $p^* > 1/2$ (resp. $p^* < 1/2$) if and only if $p > 1/2$ (resp. $p < 1/2$). The next theorem shows that the scoring step yields a consistent estimator of p^*.

Theorem 3 (Sentiment Accuracy of a New Article)**.** *Consider Models* (1)-(3)*, where Assumptions 1-4 hold,* (13) *is satisfied, and additionally,* $|T_j| \geq c_1 F_j$*, for all* $j \in S$ *and a constant* $c_1 \in (0,1)$*. Suppose* $\rho_0 \leq \widehat{\rho} \leq 1$*, for a constant* $\rho_0 \in (0,1)$*. Let* $d \in \mathbb{R}_+^m$ *denote the vector of word counts in a new article with sentiment* p*, and let* s *denote the total count of sentiment-charged words in this article. Suppose* $|p - \frac{1}{2}| \leq \rho_0 |\frac{1}{2} - c_2|$*, for a constant* $c_2 \in (0, 1/2)$*. Write*

$$err_n = \frac{1}{\sqrt{\Theta}} \left(\frac{\sqrt{|S| \log(m)}}{\sqrt{n\bar{s}\Theta}} + \frac{1}{\sqrt{s}} \right), \qquad \textit{where} \quad \Theta = \sum_{j \in S} \frac{(O_{+,j} - O_{-,j})^2}{O_{+,j} + O_{-,j}}. \tag{17}$$

We assume $s\Theta \to \infty$ *and* $err_n \to 0$*. Let* $\widehat{p}$ *be the penalized MLE in* (11)*. As* $n \to \infty$*, for each fixed* $\epsilon \in (0,1)$*, there exists a constant* $C > 0$*, such that, with probability* $1 - \epsilon$*,*

$$|\widehat{p} - p^*| \leq C \min\Big\{1, \frac{\Theta}{\lambda}\Big\} \times err_n + C \min\Big\{1, \frac{\lambda}{\Theta}\Big\} \times |p^* - \frac{1}{2}|. \tag{18}$$

Furthermore, if we set $\lambda = \frac{\Theta}{|p^* - \frac{1}{2}|} err_n$*, then* (18) *becomes*

$$|\widehat{p} - p^*| \leq C \min\big\{err_n,\ |p^* - 1/2|\big\} \leq C err_n.$$

We make a few remarks about Theorem 3. First, the term err_n in (17) accounts for the errors from the respective training step and the scoring step:

$$\frac{\sqrt{|S| \log(m)}}{\Theta \sqrt{n\bar{s}}} \qquad \text{and} \qquad \frac{1}{\sqrt{s\Theta}}.$$

Since n is large, it is the latter that will dominate in finite sample. Therefore, to guarantee $err_n \to 0$, we need $s\Theta \to \infty$.

Secondly, the choice of the regularization parameter λ reflects a bias-variance trade-off. According to (18), the first term $\min\big\{1, \frac{\Theta}{\lambda}\big\} \times err_n$ is related to the "variance", which decreases with λ; the second term $\min\big\{1, \frac{\lambda}{\Theta}\big\} \times |p^* - \frac{1}{2}|$ is the "bias", which increases with λ. In practice, it appears that most articles have a neutral tone, so that the bias is negligible relative to the variance. Besides, text data are very noisy, so imposing a large penalty in MLE significantly reduces the variance. Our estimator shares the same spirit as the James-Stein estimator (James and Stein, 1961). With the optimal choice of λ, the prediction error is bounded by the minimum of err_n and $|p^* - 1/2|$.

Thirdly, the scoring step estimates the rescaled sentiment p^* instead of p. Same as what we have explained before, this is due to that we do not have consistent estimates of $\{p_i\}_{i=1}^n$. Since $\widehat{\rho} \leq 1$, the true sentiment p is actually closer to $1/2$ than the rescaled sentiment p^*. Therefore, shrinking the sentiment towards $1/2$ helps reduce the bias. This means that although estimating p_i in the training stage creates a bias, it could be alleviated by the shrinkage effect in the scoring step. Moreover, our empirical analysis in Section 4 does not require consistent estimates of "absolute" sentiment scores. Instead, we only need "relative" sentiment scores to rank articles. It turns out sufficient to use $\widehat{p}$ for this purpose. We demonstrate how we achieve the rank consistency in Theorem 4 below.

Theorem 4 (Consistency of Rank Correlation)**.** *Under conditions of Theorem 3, suppose we are given L new articles whose sentiments $p_1, \ldots, p_L$ are i.i.d. drawn from a distribution on $[\frac{1}{2} - \rho_0(\frac{1}{2} - c_2), \frac{1}{2} + \rho_0(\frac{1}{2} - c_2)]$ with a continuous probability density. Let s_i be the count of sentiment-charged words in a new article i. We assume $C^{-1}s \leq s_i \leq Cs$, for all $1 \leq i \leq L$, where s satisfies $s\Theta/\sqrt{\log(L)} \to \infty$. We apply the estimator (11) with $\lambda \asymp \Theta err_n$ to score all new articles. Let $SR(\hat{p}, p)$ be the Spearman's rank correlation between $\{\hat{p}_i\}_{i=1}^L$ and $\{p_i\}_{i=1}^L$. As $n, m, L \to \infty$,*

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

In Theorem 3 and Theorem 4, we impose an additional condition on $O_{\pm}$, that is, $|T_j| \leq (1-c_1)F_j$ for all $j \in S$. It guarantees that the objective in (11) is strongly concave in the open set $(0, 1)$. This condition can be replaced by a restriction that $p \in [\beta, 1-\beta]$ for some constant $\beta \in (0, c_2)$.

6 Conclusion

We propose and analyze a new text-mining methodology, SESTM, 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 SESTM 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 information content of *Dow Jones Newswires* in the practical problem of portfolio construction. In this setting, our model selects intuitive lists of positive and negative words that gauge document sentiment. The resulting news sentiment scores are powerful predictors of price responses to new information. To quantify the 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 supervises 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.

Arora, Sanjeev, Rong Ge, and Ankur Moitra, 2012, Learning topic models–going beyond svd, in *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, 1–10, IEEE.

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

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.

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

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 Kelly, and Matt Taddy, 2019, Text as data, *Journal of Economic Literature* 57, 535–74.

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

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.

Janson, Svante, 2004, Large deviations for sums of partly dependent random variables, *Random Structures & Algorithms* 24, 234–248.

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.

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.

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

Mcauliffe, Jon D, and David M Blei, 2008, Supervised topic models, in *Advances in neural information processing systems*, 121–128.

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, 2014, Information transmission in finance, *Annu. Rev. Financ. Econ.* 6, 365–384.

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.

Internet Appendix

A Algorithms

Algorithm 1.

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

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

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 word j appears.

Algorithm 2.

S1. Sort the returns $\{y_i\}_{i=1}^n$ in 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 $\widehat{D} = [\widehat{d}_1, \widehat{d}_2, \ldots, \widehat{d}_n]$. 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. We also use $\widehat{O}_\pm$ to denote the two columns of $\widehat{O} = [\widehat{O}_+, \widehat{O}_-]$.

Algorithm 3.

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

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

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 Mathematical Proofs

B.1 Proof of Lemma 1

Proof. For a word j, let L_j^+ denote the total count of this word in all articles with positive returns; define L_j^- similarly. Write $t_i = \mathrm{sgn}(y_i) \in \{\pm 1\}$, for $1 \le i \le n$, and let $\bar{t} = n^{-1}\sum_{i=1}^n t_i$. It is seen that

$$L_j^{\pm} = \sum_{i=1}^n d_{j,i} \cdot (1 \pm t_i)/2, \qquad \text{and} \qquad \widehat{\pi} = (1+\bar{t})/2.$$

By definition, $f_j = L_j^+/(L_j^+ + L_j^-) = [1 + (L_j^+ - L_j^-)/(L_j^+ + L_j^-)]/2$. As a result,

$$f_j - \widehat{\pi} = \frac{1}{2}\Big(\frac{L_j^+ - L_j^-}{L_j^+ + L_j^-} - \bar{t}\Big) = \frac{\sum_{i=1}^n (t_i - \bar{t}) \cdot d_{j,i}}{\sum_{i=1}^n d_{j,i}}. \tag{B.3}$$

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

First, consider $j \in S$. As in (4), we let $F = \frac{1}{2}(O_+ + O_-)$ and $T = \frac{1}{2}(O_+ - O_-)$. We also introduce the notation $\eta_i = 2p_i - 1$. Then, $p_i O_+ + (1-p_i)O_- = F + \eta_i T$. It follows from Model (3) that

$$d_{j,i} \sim \mathrm{Binomial}\big(s_i,\ F_j + \eta_i T_j\big).$$

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

$$f_j - \widehat{\pi} \overset{(d)}{=} \frac{\sum_{i=1}^n \sum_{\ell=1}^{s_i} (t_i - \bar{t}) \cdot b_{j,i,\ell}}{\sum_{i=1}^n \sum_{\ell=1}^{s_i} b_{j,i,\ell}}, \qquad \text{where} \quad b_{j,i,\ell} \sim \mathrm{Bernoulli}(F_j + \eta_i T_j). \tag{B.4}$$

Define $\Delta_{1j} = \sum_{i=1}^n \sum_{\ell=1}^{s_i} (t_i - \bar{t})(b_{j,i,\ell} - \mathbb{E}b_{j,i,\ell})$ and $\Delta_{2j} = \sum_{i=1}^n \sum_{\ell=1}^{s_i} (b_{j,i,\ell} - \mathbb{E}b_{j,i,\ell})$. Since $\mathbb{E}b_{i,j,\ell} = F_j + \eta_i T_j$, we can re-write (B.4) as

$$f_j - \widehat{\pi} = \frac{\sum_{i=1}^n s_i (t_i - \bar{t})(F_j + \eta_i T_j) + \Delta_{1j}}{\sum_{i=1}^n s_i (F_j + \eta_i T_j) + \Delta_{2j}}.$$

Note that $\bar{t}$ it the average of t_i. It yields that $\sum_{i=1}^n s_i (t_i - \bar{t}) F_j = F_j \sum_{i=1}^n (s_i - \bar{s})(t_i - \bar{t})$. As a result,

$$f_j - \widehat{\pi} = \frac{F_j \sum_{i=1}^n (s_i - \bar{s})(t_i - \bar{t}) + T_j \sum_{i=1}^n s_i (t_i - \bar{t})\eta_i + \Delta_{1j}}{n\bar{s}F_j + T_j \sum_{i=1}^n s_i \eta_i + \Delta_{2j}}. \tag{B.5}$$

To proceed the proof, we need two technical lemmas. One is the classical Bernstein's inequality (Shorack and Wellner, 2009):

Lemma 2 (Bernstein inequality)**.** *Suppose $X_1, \cdots, X_n$ are independent random variables such that $\mathbb{E}X_i = 0$, $|X_i| \le b$ and $\mathrm{Var}(X_i) \le \sigma_i^2$ for all i. Let $\sigma^2 = \sum_{i=1}^n \sigma_i^2$. Then, for any $t > 0$,*

$$\mathbb{P}\Big(|\sum_{i=1}^n X_i| \ge t\Big) \le 2\exp\left(-\frac{t^2/2}{\sigma^2 + bt/3}\right).$$

The other is a Hoeffding-type inequality for dependent variables. In Section 5.1, we have introduced the notion of dependency graph for describing the dependency structure in the joint distribution of a set of variables. The following lemma comes from Theorem 2.1 and equation (2.2) in Janson (2004).

Lemma 3 (Hoeffding inequality for weakly dependent variables)**.** *Suppose $X_1, \cdots, X_n$ are random variables such that $\mathbb{E}X_i = 0$ and $a_i \le X_i \le b_i$. Suppose Γ is a valid dependency graph for the joint distribution of $\{X_i\}_{i=1}^n$. The maximum degree of Γ is denoted by $d^*(\Gamma)$. Then, for any $t > 0$,*

$$\mathbb{P}\Big(|\sum_{i=1}^n X_i| \ge t\Big) \le 2\exp\left(-\frac{2t^2}{[d^*(\Gamma)+1]\sum_{i=1}^n (b_i - a_i)^2}\right).$$

When applying Lemmas 2-3, we often use particular choices of t. In Lemma 2, for any $\epsilon \in (0,1)$, we can let $t = \sigma\sqrt{2\log(2/\epsilon)} + (2b/3)\log(2/\epsilon)$ to make $\mathbb{P}(|\sum_{i=1}^n X_i| \ge t)$ bounded by ϵ. In particular, as $n \to \infty$, for any sequence $\epsilon_n \to 0$, there exists a constant $C > 0$ such that

$$|\sum_{i=1}^n X_i| \le C\big[\sigma\sqrt{\log(1/\epsilon_n)} + b\log(1/\epsilon_n)\big] \tag{B.6}$$

holds with probability $1 - \epsilon_n$. Similarly, in Lemma 3, for any sequence $\epsilon_n \to 0$, we can choose an appropriate t to prove: With probability $1 - \epsilon_n$,

$$|\sum_{i=1}^n X_i| \le C\sqrt{d^*(\Gamma)\sum_{i=1}^n (b_i - a_i)^2 \log(1/\epsilon_n)}. \tag{B.7}$$

In the proof below, we use (B.6) and (B.7), instead of the original statements of Lemmas 2-3.

With the technical preparation, we now proceed to study the right hand side of (B.5). First, we bound $|\Delta_{1j}|$ and $|\Delta_{2j}|$. The random variables $\{b_{j,i,\ell} - \mathbb{E}b_{j,i,\ell}\}$ are independent, mean-zero, satisfying that $\sum_{i=1}^n \sum_{\ell=1}^{s_i} \mathrm{var}(b_{j,i,\ell}) \le \sum_{i=1}^n \sum_{\ell=1}^{s_i} (F_j + \eta_i T_j) \le \sum_{i=1}^n \sum_{\ell=1}^{s_i} 2F_j = 2n\bar{s}F_j$. By (B.6), with probability $1 - m^{-1}\epsilon_n$,

$$\begin{aligned} |\Delta_{2j}| &= \Big|\sum_{i=1}^n \sum_{\ell=1}^{s_i} (b_{j,i,\ell} - \mathbb{E}b_{j,i,\ell})\Big| \le C\sqrt{n\bar{s}F_j \log(m/\epsilon_n)} + C\log(m/\epsilon_n) \\ &\le C\sqrt{n\bar{s}F_j \log(m/\epsilon_n)}, \end{aligned} \tag{B.8}$$

where the last line is due to the assumption $n\bar{s}F_j/\log(m/\epsilon_n) \to \infty$. Recall that $\Delta_{1j} = \sum_{i=1}^n \sum_{\ell=1}^{s_i} (t_i - \bar{t})(b_{j,i,\ell} - \mathbb{E}b_{j,i,\ell})$. We condition on $\{y_i\}_{i=1}^n$. Then, t_i's are non-random; additionally, by Assumption 3, $\{d_{j,i}\}_{1 \le i \le n}$ have the same Binomial distributions as in the unconditional case. By similar arguments as above, conditioning on $\{y_i\}_{i=1}^n$, with probability $1 - m^{-1}\epsilon_n$,

$$|\Delta_{1j}| \le C\sqrt{F_j \sum_{i=1}^n s_i (t_i - \bar{t})^2 \log(m/\epsilon_n)} + C(\max_i |t_i - \bar{t}|)\log(m/\epsilon_n)$$

$$\leq C\sqrt{n\bar{s}F_j \log(m/\epsilon_n)}, \tag{B.9}$$

where the last inequality is due to $|t_i - \bar{t}| \leq 1$. Since the bound in (B.9) does not depend on y_i's, the inequality also holds unconditionally with the same probability.

We then insert (B.8)-(B.9) into (B.5). Consider the term $(n\bar{s}F_j + T_j \sum_{i=1}^n s_i\eta_i)$ in the denominator of (B.5). It can be re-written as $\sum_{i=1}^n s_i[p_i O_{+,j} + (1-p_i)O_{-,j}]$, which, by item (d) in Assumption 4, is lower bounded by $c_0 \sum_{i=1}^n s_i(O_{+,j} + O_{-,j}) = 2c_0\bar{s}F_j$, for a constant $c_0 > 0$. At the same time, since $|\eta_i| \leq 1$ and $T_j \leq F_j$, this term also has an upper bound: $n\bar{s}F_j + T_j \sum_{i=1}^n s_i\eta_i \leq 2n\bar{s}F_j$. Combining the above results with (B.5) gives

$$\begin{aligned} |f_j - \widehat{\pi}| &\geq \frac{|T_j \sum_{i=1}^n s_i(t_i - \bar{t})\eta_i|}{2n\bar{s}F_j + |\Delta_{2j}|} - \frac{F_j|\sum_{i=1}^n (s_i - \bar{s})(t_i - \bar{t})|}{2c_0 n\bar{s}F_j - |\Delta_{2j}|} - \frac{|\Delta_{1j}|}{2c_0 n\bar{s}F_j - |\Delta_{2j}|} \\ &\gtrsim \frac{|T_j \sum_{i=1}^n s_i(t_i - \bar{t})\eta_i|}{2n\bar{s}F_j} - O\bigg(\frac{|\sum_{i=1}^n (s_i - \bar{s})(t_i - \bar{t})|}{n\bar{s}} + \frac{\sqrt{\log(m/\epsilon_n)}}{\sqrt{n\bar{s}F_j}}\bigg). \end{aligned} \tag{B.10}$$

In the second line, we have plugged in the bounds in (B.8)-(B.9); furthermore, we have used the fact that $|\Delta_{1j}| = o(n\bar{s}F_j)$ and $|\Delta_{2j}| = o(n\bar{s}F_j)$, due to $n\bar{s}F_j/\log(m/\epsilon_n) \to \infty$.

Last, we deal with the randomness of $\{t_i\}_{i=1}^n$ in (B.10). Recall that $t_i = \mathrm{sgn}(y_i)$. By Assumption 3, y_i's are dependent, and the dependence structure is captured by a dependency graph with maximum degree K_n. It is easy to see that the same graph can be used as a dependency graph for t_i's. Hence, we can apply (B.7) with $d^*(\Gamma) = K_n$.

In (B.10), we first study the term $|\sum_{i=1}^n s_i(t_i - \bar{t})\eta_i|$. Note that $\mathbb{E}t_i = 2g(p_i) - 1$ and $\mathbb{E}\bar{t} = 2\bar{g} - 1$. We thus have the decomposition $t_i - \bar{t} = 2[g(p_i) - \bar{g}] + (t_i - \mathbb{E}t_i) - (\bar{t} - \mathbb{E}\bar{t})$. It follows that

$$\begin{aligned} \Big|\sum_{i=1}^n s_i(t_i - \bar{t})\eta_i\Big| &\geq 2\Big|\sum_{i=1}^n s_i\eta_i[g(p_i) - \bar{g}]\Big| - \Big|\sum_{i=1}^n s_i\eta_i(t_i - \mathbb{E}t_i)\Big| - \Big|\sum_{i=1}^n s_i\eta_i(\bar{t} - \mathbb{E}\bar{t})\Big| \\ &= 4n\bar{s}\theta - \Big|\sum_{i=1}^n s_i\eta_i(t_i - \mathbb{E}t_i)\Big| - \Big|n^{-1}\sum_{i=1}^n s_i\eta_i\Big|\Big|\sum_{i=1}^n (t_i - \mathbb{E}t_i)\Big|, \end{aligned} \tag{B.11}$$

where in the last line we have used the definition of θ in (12) and the equality $\bar{t} - \mathbb{E}\bar{t} = n^{-1}\sum_{i=1}^n (t_i - \mathbb{E}t_i)$. We apply (B.7) to $X_i = t_i - \mathbb{E}t_i$. Since $|t_i - \mathbb{E}t_i| \leq 1$, it gives that, with probability $1 - m^{-1}\epsilon_n$,

$$\Big|\sum_{i=1}^n (t_i - \mathbb{E}t_i)\Big| \leq C\sqrt{nK_n \log(m/\epsilon_n)}.$$

Similarly, we apply (B.7) to $X_i = s_i\eta_i(t_i - \mathbb{E}t_i)$. Since $|\eta_i| = |2p_i - 1| \leq 1$ and $s_i \leq C\bar{s}$ (by item (a) of Assumption 4), it follows that, with probability $1 - m^{-1}\epsilon_n$,

$$\Big|\sum_{i=1}^n s_i\eta_i(t_i - \mathbb{E}t_i)\Big| \leq C\bar{s}\sqrt{nK_n \log(m/\epsilon_n)}.$$

Plugging the above results into (B.11) and noting that $|n^{-1}\sum_{i=1}^n s_i\eta_i| \leq C\bar{s}$, we immediately have:

With probability $1 - m^{-1}\epsilon_n$,

$$\Big|\sum_{i=1}^n s_i(t_i - \bar{t})\eta_i\Big| \geq 4n\bar{s}\theta - O\Big(\bar{s}\sqrt{nK_n \log(m/\epsilon_n)}\Big). \tag{B.12}$$

Next, we study the term $|\sum_{i=1}^n (s_i - \bar{s})(t_i - \bar{t})|$ in (B.10). Using the decomposition $t_i - \bar{t} = 2[g(p_i) - \bar{g}] + (t_i - \mathbb{E}t_i) - (\bar{t} - \mathbb{E}\bar{t})$ again, we get:

$$\Big|\sum_{i=1}^n (s_i - \bar{s})(t_i - \bar{t})\Big| \leq 2\Big|\sum_{i=1}^n (s_i - \bar{s})[g(p_i) - \bar{g}]\Big| + \Big|\sum_{i=1}^n (s_i - \bar{s})(t_i - \mathbb{E}t_i)\Big| + \Big|\sum_{i=1}^n (s_i - \bar{s})(\bar{t} - \mathbb{E}\bar{t})\Big|.$$

The last term is zero. The first term is bounded by $C\bar{s}\sqrt{nK_n \log(m)}$, due to item (e) of Assumption 4. The second term can be bounded using (B.7) again. We omit the details and state the results directly: With probability $1 - m^{-1}\epsilon_n$, $|\sum_{i=1}^n (s_i - \bar{s})(t_i - \mathbb{E}t_i)| \leq C\bar{s}\sqrt{nK_n \log(m/\epsilon_n)}$. Combining these results gives

$$\Big|\sum_{i=1}^n (s_i - \bar{s})(t_i - \bar{t})\Big| \leq C\bar{s}\sqrt{nK_n \log(m/\epsilon_n)}. \tag{B.13}$$

We plug (B.12) and (B.13) into (B.10). It follows that, with probability $1 - m^{-1}\epsilon_n$,

$$|f_j - \widehat{\pi}| \gtrsim \frac{2\theta|T_j|}{F_j} - O\bigg(\frac{\sqrt{K_n \log(m/\epsilon_n)}}{\sqrt{n}} + \frac{\sqrt{\log(m/\epsilon_n)}}{\sqrt{n\bar{s}F_j}}\bigg).$$

Combined with the probability union bound, it yields that the above inequality holds simultaneously for all $j \in S$ with probability $1 - \epsilon_n$. This proves the first claim.

Next, consider $j \in N$. Note that $f_j - \widehat{\pi}$ has the same expression as in (B.3), and we aim to study its numerator and denominator. By Assumption 3,

$$d_{j,i} \,|\, \{y_1, \ldots, y_n\} \sim \text{Binomial}\big(n_i,\, \Omega_{j,i}(y_i)\big), \qquad 1 \leq i \leq n.$$

Similarly as before, we introduce random variables $\{h_{j,i,\ell}\}_{1\leq i\leq n, 1\leq \ell\leq n_i}$: conditioning on $\{y_1, \ldots, y_n\}$, they are independent, and $h_{j,i,\ell} \sim \text{Bernoulli}(\Omega_{j,i}(y_i))$. Then, $d_{j,i}$ can be replaced by $\sum_{\ell=1}^{n_i} h_{j,i,\ell}$. We thus re-write the numerator and denominator in (B.3) as

$$U_j \equiv \sum_{i=1}^n (t_i - \bar{t})d_{j,i} = \sum_{i=1}^n \sum_{\ell=1}^{n_i} (t_i - \bar{t})h_{j,i,\ell}, \qquad \text{and} \qquad V_j \equiv \sum_{i=1}^n d_{j,i} = \sum_{i=1}^n \sum_{\ell=1}^{n_i} h_{j,i,\ell}.$$

We first study U_j. Re-write

$$U_j = \sum_{i=1}^n \sum_{\ell=1}^{n_i} (t_i - \bar{t})\big[h_{j,i,\ell} - \Omega_{j,i}(y_i)\big] + \sum_{i=1}^n n_i(t_i - \bar{t})\Omega_{j,i}(y_i) \equiv (I_1) + (I_2).$$

Conditioning on $\{y_i\}_{i=1}^n$, the variables $\{(t_i - \bar{t})(h_{j,i,\ell} - \Omega(y_i))\}_{1\leq i\leq n, 1\leq \ell\leq n_i}$ are mutually independent, $|h_{j,i,\ell} - \Omega_{j,i}(y_i)| \leq 1$, and $\sum_{i=1}^n \sum_{\ell=1}^{n_i} \text{var}(h_{j,i,\ell}|y_1, \ldots, y_n) \leq \sum_{i=1}^n n_i\Omega_{j,i}(y_i) \leq Cq_j \sum_{i=1}^n n_i = Cn\bar{n}q_j$,

where we have used item (b) of Assumption 4. We apply (B.6) to get: Conditioning on $\{y_i\}_{i=1}^n$, with probability $1 - m^{-1}\epsilon_n$,

$$|(I_1)| \le C\sqrt{n\bar{n}q_j \log(m/\epsilon_n)} + C\log(m/\epsilon_n) \le C\sqrt{n\bar{n}q_j \log(m/\epsilon_n)}, \tag{B.14}$$

where the last inequality comes from $n\bar{n}q_j/\log(m/\epsilon_n) \to \infty$. This bound does not depend on y_i's, therefore, it also holds marginally without conditioning. We then bound $|(I_2)|$. Write

$$(I_2) = Z - \Delta, \qquad \text{where} \quad Z = \sum_{i=1}^n n_i(t_i - \mathbb{E}\bar{t})\,\Omega_{j,i}(y_i), \quad \Delta = \sum_{i=1}^n n_i(\bar{t} - \mathbb{E}\bar{t})\,\Omega_{j,i}(y_i).$$

We use Assumption 1. Since $\mathbb{E}[(t_i - \bar{t})\Omega_{j,i}(y_i)] = \mathbb{E}\big[(t_i - \bar{t}) \cdot \mathbb{E}(n_i^{-1} d_{j,i} | y_1, \ldots, y_n)\big] = \mathbb{E}[n_i^{-1}(t_i - \bar{t})d_{j,i}]$, it holds that

$$\mathbb{E}[(I_2)] = \sum_{i=1}^n \mathbb{E}[(\mathrm{sgn}(y_i) - \overline{\mathrm{sgn}}_y)\, d_{j,i}] = 0. \tag{B.15}$$

As a result,

$$|(I_2)| = \big|(I_2) - \mathbb{E}[(I_2)]\big| \le |Z - \mathbb{E}Z| + |\Delta| + \mathbb{E}|\Delta|. \tag{B.16}$$

We bound the three terms on the right hand separately. By item (b) of Assumption 4, it is always true that $0 \le \Omega_{j,i}(y_i) \le Cq_j$. As a result, $|\Delta| = |\bar{t} - \mathbb{E}\bar{t}| \cdot |\sum_{i=1}^n n_i \Omega_{j,i}(y_i)| \le |\bar{t} - \mathbb{E}\bar{t}| \cdot Cn\bar{n}q_j \le C\bar{n}q_j |\sum_{i=1}^n (t_i - \mathbb{E}t_i)|$. In the equation below (B.11), we have already proved that $|\sum_{i=1}^n (t_i - \mathbb{E}t_i)| \le C\sqrt{nK_n \log(m/\epsilon_n)}$, with probability $1 - m^{-1}\epsilon_n$. Combining them gives

$$|\Delta| \le C\bar{n}q_j \sqrt{nK_n \log(m/\epsilon_n)}.$$

Furthermore, if we apply Lemma 3, instead of (B.7), to $\sum_{i=1}^n (t_i - \mathbb{E}t_i)$, we will find out that this is a sub-Gaussian variable with a sub-Gaussian norm $O(nK_n)$. In particular, it is first absolute moment is $O(\sqrt{nK_n})$. It follows that

$$\mathbb{E}|\Delta| \le C\bar{n}q_j\, \mathbb{E}\Big(\Big|\sum_{i=1}^n (t_i - \mathbb{E}t_i)\Big|\Big) \le C\bar{n}q_j \sqrt{nK_n}.$$

Write $Z - \mathbb{E}Z = \sum_{i=1}^n n_i\big\{(t_i - \mathbb{E}\bar{t})\Omega_{j,i}(y_i) - \mathbb{E}\big[(t_i - \mathbb{E}\bar{t})\Omega_{j,i}(y_i)\big]\big\}$. We shall apply (B.7). Note that each summand only depends on y_i. Hence, the dependency graph in Assumption 3 is still a valid dependency graph here. Additionally, each summand is upper bounded by $n_i \cdot Cq_j|t_i - \mathbb{E}\bar{t}| \le C\bar{n}q_j$, where we have used item (a) of Assumption 4 which says $n_{\max} \le C\bar{n}$. It follows from (B.7) that, with probability $1 - m^{-1}\epsilon_n$,

$$|Z - \mathbb{E}Z| \le C\bar{n}q_j \sqrt{nK_n \log(m/\epsilon_n)}.$$

We plug the above inequalities into (B.16) and find out that

$$|(I_2)| \le C\bar{n}q_j \sqrt{nK_n \log(m/\epsilon_n)}. \tag{B.17}$$

Combining (B.14) and (B.17) gives

$$|U_j| \le C\Big(\sqrt{n\bar{n}q_j \log(m/\epsilon_n)} + \bar{n}q_j\sqrt{nK_n \log(m/\epsilon_n)}\Big). \tag{B.18}$$

We then study V_j. Rewrite

$$V_j = \sum_{i=1}^{n} n_i\, \mathbb{E}[\Omega_{j,i}(y_i)] + \sum_{i=1}^{n} n_i\big\{\Omega_{j,i}(y_i) - \mathbb{E}[\Omega_{j,i}(y_i)]\big\} + \sum_{i=1}^{n}\sum_{\ell=1}^{n_i}\big[h_{j,i,\ell} - \Omega_{j,i}(y_i)\big].$$

The last two terms can be bounded in similar ways as we bound $|Z - \mathbb{E}Z|$ and $|(I_1)|$. Moreover, using items (a) and (b) of Assumption 4, we have $\sum_{i=1}^{n} n_i\, \mathbb{E}[\Omega_{j,i}(y_i)] \ge n_{\min}\sum_{i=1}^{n}\mathbb{E}[\Omega_{j,i}(y_i)] = nn_{\min}q_j \ge C^{-1}n\bar{n}q_j$. It follows that

$$\begin{aligned} V_j &\ge C^{-1}n\bar{n}q_j - C\Big(\sqrt{n\bar{n}q_j\log(m/\epsilon_n)} + \bar{n}q_j\sqrt{nK_n\log(m/\epsilon_n)}\Big) \\ &\ge C^{-1}n\bar{n}q_j - o\big(n\bar{n}q_j\big), \end{aligned} \tag{B.19}$$

where the last inequality is due to $n\bar{n}q_j/\log(m/\epsilon_n) \to \infty$ and $nK_n^{-1}/\log(m/\epsilon_n) \to \infty$. We put (B.18) and (B.19) together to get: With probability $1 - m^{-1}\epsilon_n$,

$$|f_j - \widehat{\pi}| = \frac{|U_j|}{V_j} = O\bigg(\frac{\sqrt{\log(m/\epsilon_n)}}{\sqrt{n\bar{n}q_j}} + \frac{\sqrt{K_n\log(m/\epsilon_n)}}{\sqrt{n}}\bigg).$$

This proves the second claim. □

B.2 Proof of Theorem 1

Proof. In Lemma 1, letting $\epsilon_n = 1/m$, we have: With probability $1 - O(m^{-1})$, simultaneously for all $1 \le j \le m$,

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

where $e_n^2 = (\min\{K_n^{-1},\ \bar{s}\min_{j\in S} F_j,\ \bar{n}\min_{j\in N} q_j\})^{-1}\frac{\log(m)}{n}$. The assumption (13) ensures $\theta F_j^{-1}|T_j| \gg e_n\sqrt{\log(m)}$. 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. This proves the claim. □

B.3 Proof of Theorem 2

Proof. By Theorem 1, $\mathbb{P}(\widehat{S} = S) = 1 - o(1)$. Hence, it suffices to prove the claim by replacing $\widehat{S}$ with S in the estimation step. In particular, we replace $\widehat{h}_i$ in (8) by $h_i = d_{[S],i}/s_i$. Write $H = [h_1, \ldots, h_n]$. The estimate $\widehat{O}$ is obtained by modifying and renormalizing

$$\widetilde{O} = H\widehat{W}'(\widehat{W}\widehat{W}')^{-1}.$$

By Model (3), $\mathbb{E}H = OW$. It motivates us to define an intermediate matrix

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

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

$$\|F^* - F\|_1 = O(n^{-1}), \qquad \|T^* - \widehat{\rho}T\|_1 = O(n^{-1}). \tag{B.20}$$

In the second part of our proof, we show that

$$\|\widehat{O}_\pm - O^*_\pm\|_1 \leq C\sqrt{|S|\log(m)/(n\bar{s})} + O(n^{-1}). \tag{B.21}$$

The claim follows by combining (B.20)-(B.21) and noting that $\max\{\|\widehat{F}-F^*\|_1, \|\widehat{T}-T^*\|_1\} \leq \frac{1}{2}(\|\widehat{O}_+ - O^*_+\|_1 + \|\widehat{O}_- - O^*_-\|_1)$.

First, we show (B.20). By definition,

$$\begin{aligned}[F^*,\ T^*] = O^*\begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & -\frac{1}{2}\end{bmatrix} &= O(W\widehat{W}')(\widehat{W}\widehat{W}')^{-1}\begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & -\frac{1}{2}\end{bmatrix}\\ &= [F,\ T]\underbrace{\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}(W\widehat{W}')(\widehat{W}\widehat{W}')^{-1}\begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & -\frac{1}{2}\end{bmatrix}}_{\equiv M}.\end{aligned} \tag{B.22}$$

We now calculate the 2×2 matrix M. With the returns sorted in the ascending order, $y_{(1)} < y_{(2)} < \ldots < y_{(n)}$, the estimation step (9) sets $\widehat{p}_{(i)} = i/n$, for $1 \leq i \leq n$. 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

$$n^{-1}(\widehat{W}\widehat{W}') = \begin{bmatrix}1/3 & 1/6\\ 1/6 & 1/3\end{bmatrix} + O(n^{-1}) \quad\Longrightarrow\quad n(\widehat{W}\widehat{W}')^{-1} = \begin{bmatrix}4 & -2\\ -2 & 4\end{bmatrix} + O(n^{-1}), \tag{B.23}$$

where $O(n^{-1})$ is with respect to the element-wise maximum norm $\|\cdot\|_{\max}$ of a matrix. Additionally, the definition of $\widehat{W}$ in (10) yields

$$n^{-1}(W\widehat{W}') = \begin{bmatrix}\frac{1}{n}\sum_i p_i\widehat{p}_i & \frac{1}{n}\sum_i p_i(1-\widehat{p}_i)\\ \frac{1}{n}\sum_i(1-p_i)\widehat{p}_i & \frac{1}{n}\sum_i(1-p_i)(1-\widehat{p}_i)\end{bmatrix}. \tag{B.24}$$

We now plug (B.23)-(B.24) into (B.22). It gives

$$\begin{aligned} M &= \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{n}\sum_i p_i\widehat{p}_i & \frac{1}{n}\sum_i p_i(1-\widehat{p}_i) \\ \frac{1}{n}\sum_i (1-p_i)\widehat{p}_i & \frac{1}{n}\sum_i (1-p_i)(1-\widehat{p}_i) \end{bmatrix} \cdot \begin{bmatrix} 4 & -2 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} + O(n^{-1}) \\ &= \begin{bmatrix} \frac{1}{n}\sum_i \widehat{p}_i & \frac{1}{n}\sum_i (1-\widehat{p}_i) \\ \frac{1}{n}\sum_i (2p_i-1)\widehat{p}_i & \frac{1}{n}\sum_i (2p_i-1)(1-\widehat{p}_i) \end{bmatrix} \cdot \begin{bmatrix} 1 & 3 \\ 1 & -3 \end{bmatrix} + O(n^{-1}) \\ &= \begin{bmatrix} 1 & \frac{6}{n}\sum_i (\widehat{p}_i - \frac{1}{2}) \\ \frac{2}{n}\sum_i (p_i - \frac{1}{2}) & \frac{12}{n}\sum_i (p_i-\frac{1}{2})(\widehat{p}_i - \frac{1}{2}) \end{bmatrix} + O(n^{-1}). \end{aligned}$$

By Assumption 2, we have $\sum_i (p_i - 1/2) = 0$. Additionally, the way we construct $\{\widehat{p}_i\}_{i=1}^n$ guarantees $\sum_i (\widehat{p}_i - 1/2) = 0$. Combining these results and using the definition of $\widehat{\rho}$ in (15), we have

$$M = \begin{bmatrix} 1 & 0 \\ 0 & \widehat{\rho} \end{bmatrix} + \Delta, \qquad \text{where} \quad \|\Delta\|_{\max} = O(n^{-1}). \tag{B.25}$$

We plug (B.25) into (B.22). It gives

$$F^* = F + [F,\, T]\Delta, \qquad T^* = \widehat{\rho} T + [F,\, T]\Delta.$$

Let $\|\cdot\|_1$ denote the matrix L_1-norm, which is equal to the maximum column sum of the matrix. It is easy to see that $\|[F,T]\|_1 \le 1$. Moreover, $\|\Delta\|_1 \le 2\|\Delta\|_{\max} = O(n^{-1})$. It follows that

$$\|F^* - F\|_1 \le \|[F,T]\|_1 \|\Delta\|_1 = O(n^{-1}), \qquad \|T^* - \widehat{\rho} T\|_1 \le \|[F,T]\|_1 \|\Delta\|_1 = O(n^{-1}).$$

This proves (B.20)

Second, we show (B.21). Write $O_{\pm}^{(\widehat{\rho})} = F \pm \widehat{\rho} T$. Note that F is a nonnegative vector whose entries sum to 1, T is a vector whose entries sum to 0, and $|T_j| < F_j$ for each j. As a result, when $0 < \widehat{\rho} \le 1$, the two vectors $O_{\pm}^{(\widehat{\rho})}$ are nonnegative and satisfy $\|O_{\pm}^{(\widehat{\rho})}\|_1 = 1$. Also, it follows from (B.20) that

$$\|O_{\pm}^* - O_{\pm}^{(\widehat{\rho})}\|_1 = O(n^{-1}). \tag{B.26}$$

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

$$|\widehat{O}_{\pm,j} - O_{\pm,j}^{(\widehat{\rho})}| \le |\overline{O}_{\pm,j} - O_{\pm,j}^{(\widehat{\rho})}| + |\overline{O}_{\pm,j}| \cdot \Big| \frac{1}{\|\overline{O}_{\pm}\|_1} - 1 \Big|.$$

Since $\|O_{\pm}^{(\widehat{\rho})}\|_1 = 1$, it holds that $|\|\overline{O}_{\pm}\|_1^{-1} - 1| = \|\overline{O}_{\pm}\|_1^{-1} |\|\overline{O}_{\pm}\|_1 - \|O_{\pm}^{(\widehat{\rho})}\|_1| \le \|\overline{O}_{\pm}\|_1^{-1} \|\overline{O}_{\pm} - O_{\pm}^{(\widehat{\rho})}\|_1$. Hence,

$$|\widehat{O}_{\pm,j} - O_{\pm,j}^{(\widehat{\rho})}| \le |\overline{O}_{\pm,j} - O_{\pm,j}^{(\widehat{\rho})}| + \frac{|\overline{O}_{\pm,j}|}{\|\overline{O}_{\pm}\|_1} \|\overline{O}_{\pm} - O_{\pm}^{(\widehat{\rho})}\|_1. \tag{B.27}$$

Summing over j on both sides gives

$$\|\widehat{O}_\pm - O_\pm^{(\widehat{\rho})}\|_1 \le 2\|\overline{O}_\pm - O_\pm^{(\widehat{\rho})}\|_1.$$

Moreover, since $O_\pm^{(\widehat{\rho})}$ are nonnegative vectors, the operation of truncating out negative entries in $\widetilde{O}_\pm$ (to obtain $\overline{O}_\pm$) always makes it closer to $O_\pm^{(\widehat{\rho})}$. It implies $\|\overline{O}_\pm - O_\pm^{(\widehat{\rho})}\|_1 \le \|\widetilde{O}_\pm - O_\pm^{(\widehat{\rho})}\|_1$. Combining the above gives

$$\|\widehat{O}_\pm - O_\pm^{(\widehat{\rho})}\|_1 \le 2\|\widetilde{O}_\pm - O_\pm^{(\widehat{\rho})}\|_1. \tag{B.28}$$

It follows that

$$\begin{aligned}\|\widehat{O}_\pm - O_\pm^*\|_1 &\le \|\widehat{O}_\pm - O_\pm^{(\widehat{\rho})}\|_1 + \|O_\pm^{(\widehat{\rho})} - O_\pm^*\|_1\\ &\le 2\|\widetilde{O}_\pm - O_\pm^{(\widehat{\rho})}\|_1 + \|O_\pm^{(\widehat{\rho})} - O_\pm^*\|_1\\ &\le 2\|\widetilde{O}_\pm - O_\pm^*\|_1 + 3\|O_\pm^{(\widehat{\rho})} - O_\pm^*\|_1\\ &\le 2\|\widetilde{O}_\pm - O_\pm^*\|_1 + O(n^{-1}),\end{aligned}$$

where the second line is from (B.28) and the last line is due to (B.26). Therefore, to show (B.21), it suffices to show that

$$\|\widetilde{O}_\pm - O_\pm^*\|_1 \le C\sqrt{|S|\log(m)/(n\bar{s})}. \tag{B.29}$$

We now show (B.29). As mentioned before, $\mathbb{E}H = OW$. Define $Z = H - OW$ and write

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

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

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

It follows that

$$\begin{aligned}|\widetilde{O}_{\pm,j} - O^*_{\pm,j}| &\le \big\|[Z\widehat{W}'(\widehat{W}\widehat{W}')^{-1}]_{j,\cdot}\big\|_\infty\\ &\le \bigg(\Big|\frac{1}{n}\sum_{i=1}^n \widehat{p}_i Z_{j,i}\Big| + \Big|\frac{1}{n}\sum_{i=1}^n (1-\widehat{p}_i) Z_{j,i}\Big|\bigg)\big\|n(\widehat{W}\widehat{W}')^{-1}\big\|_\infty\\ &\le C\bigg(\Big|\frac{1}{n}\sum_{i=1}^n \widehat{p}_i Z_{j,i}\Big| + \Big|\frac{1}{n}\sum_{i=1}^n (1-\widehat{p}_i) Z_{j,i}\Big|\bigg),\end{aligned} \tag{B.30}$$

where in the last line we have used (B.23). We now bound $|\frac{1}{n}\sum_{i=1}^n \widehat{p}_i Z_{j,i}|$. The bound for $|\frac{1}{n}\sum_{i=1}^n (1-\widehat{p}_i) Z_{j,i}|$ can be obtained similarly, so the proof is omitted. By definition, $Z_{j,i} = H_{j,i} - \mathbb{E}H_{j,i}$, where $H_{j,i} = s_i^{-1} d_{j,i}$. By Assumption 3, conditioning on $\{y_i\}_{i=1}^n$, the distribution of $d_{[S],i}$ is a multinomial distribution. It follows that

$$d_{j,i}\big|\{y_1, y_2, \ldots, y_n\} \sim \text{Binomial}\big(s_i,\ p_i O_{+,j} + (1-p_i) O_{-,j}\big).$$

Let $\{b_{j,i,\ell}\}_{\ell=1}^{s_i}$ be a collection of iid Bernoulli variables with a success probability $[p_i O_{+,j}+(1-p_i)O_{-,j}]$. Then, $d_{j,i}$ has the same distribution as $\sum_{\ell=1}^{s_i} b_{j,i,\ell}$. As a result, $Z_{j,i} \overset{(d)}{=} \sum_{\ell=1}^{s_i} s_i^{-1}(b_{j,i,\ell} - \mathbb{E}b_{j,i,\ell})$, and we can write

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

Conditioning on $\{y_i\}_{i=1}^n$, $\widehat{p}_i$'s now become non-random. We shall apply (B.6). Note that the variables $\widehat{p}_i s_i^{-1}(b_{j,i,\ell} - \mathbb{E}b_{j,i,\ell})$ 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 (B.6), with probability $1-O(m^{-2})$,

$$\Big|\sum_{i=1}^{n} \widehat{p}_i Z_{j,i}\Big| \le C\sqrt{n\bar{s}^{-1}F_j\log(m)} + C\bar{s}^{-1}\log(m) \le C\sqrt{n\bar{s}^{-1}F_j\log(m)}, \tag{B.31}$$

where the last line is due to $n\bar{s}F_j/\log(m) \to \infty$. The bound for $|\sum_{i=1}^n (1-\widehat{p}_i)Z_{j,i}|$ is similar. Plugging them into (B.30) gives

$$|\widetilde{O}_{\pm,j} - O^*_{\pm,j}| \le C\frac{\sqrt{F_j\log(m)}}{\sqrt{n\bar{s}}}. \tag{B.32}$$

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 (B.21). The proof is now complete. □

B.4 Proof of Corollary 1

Proof. Define $O^* = OW\widehat{W}'(\widehat{W}\widehat{W}')^{-1}$ similarly as in the proof of Theorem 2, except that $\widehat{W}$ is now constructed from an arbitrary estimate $\widehat{p}_i$. In the proof of Theorem 2, we have shown $\|\widehat{O}_{\pm} - O^*_{\pm}\|_1 \le C\sqrt{|S|\log(m)/(n\bar{s})} + O(n^{-1})$, with probability $1-o(1)$. The proof there does not use any particular structure of $\widehat{p}_i$, so the conclusion continues to hold for an arbitrary $\widehat{p}_i$. As a result, to show the claim, it suffices to show that,

$$\max\{\|F^* - F\|_1,\ \|T^* - T\|_1\} \le C\sqrt{n^{-1}\sum_{i=1}^{n}(\widehat{p}_i - p_i)^2}. \tag{B.33}$$

Below, we prove (B.33). Note that we can re-write $O = O\widehat{W}\widehat{W}'(\widehat{W}\widehat{W}')^{-1}$. It follows that $O^* - O = O(W-\widehat{W})\widehat{W}'(\widehat{W}\widehat{W}')^{-1}$. We thus write

$$\begin{aligned}[F^* - F,\ T^* - T] &= (O^* - O)\begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & -\frac{1}{2}\end{bmatrix} = O(W-\widehat{W})\widehat{W}'(\widehat{W}\widehat{W}')^{-1}\begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & -\frac{1}{2}\end{bmatrix}\\ &= [F,\ T]\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}(W-\widehat{W})\widehat{W}'(\widehat{W}\widehat{W}')^{-1}\begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & -\frac{1}{2}\end{bmatrix}\end{aligned}$$

$$= [F,\ T] \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{n}\sum_i (p_i - \widehat{p}_i)\widehat{p}_i & \frac{1}{n}\sum_i (p_i - \widehat{p}_i)(1-\widehat{p}_i) \\ \frac{1}{n}\sum_i (\widehat{p}_i - p_i)\widehat{p}_i & \frac{1}{n}\sum_i (\widehat{p}_i - p_i)(1-\widehat{p}_i) \end{bmatrix} \cdot n(\widehat{W}\widehat{W}')^{-1} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}$$
$$= [F,\ T] \begin{bmatrix} 0 & 0 \\ \frac{2}{n}\sum_i (p_i - \widehat{p}_i)\widehat{p}_i & \frac{2}{n}\sum_i (p_i - \widehat{p}_i)(1-\widehat{p}_i) \end{bmatrix} \cdot \underbrace{n(\widehat{W}\widehat{W}')^{-1} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}}_{M^*}.$$

By our assumption, $\|n(\widehat{W}\widehat{W}')^{-1}\| \le a_0^{-1}$. Then, the 2×2 matrix M^* satisfies that $\|M^*\| \le C$. As a result,

$$\max\{\|F^* - F\|_1,\ \|T^* - T\|_1\} \le C\|[F,\ T]\|_1 \cdot \max\left\{\left|\frac{1}{n}\sum_{i=1}^n (\widehat{p}_i - p_i)\widehat{p}_i\right|,\ \left|\frac{1}{n}\sum_{i=1}^n (\widehat{p}_i - p_i)(1-\widehat{p}_i)\right|\right\}.$$

By Cauchy-Schwarz inequality, $|n^{-1}\sum_i (\widehat{p}_i - p_i)\widehat{p}_i| \le n^{-1}\sqrt{\sum_i (\widehat{p}_i - p_i)^2}\sqrt{\sum_i \widehat{p}_i^2} \le \sqrt{n^{-1}\sum_i (\widehat{p}_i - p_i)^2}$. A similar bound holds for $|n^{-1}\sum_i (\widehat{p}_i - p_i)(1-\widehat{p}_i)|$. Moreover, note that $\|[F,\ T]\|_1 = \max\{\|F\|_1, \|T\|_1\} = 1$. Plugging them into the above inequality gives (B.33). This completes the proof. □

B.5 Proof of Theorem 3

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

Recall that p and p^* are the true sentiment and rescaled sentiment of the new article, respectively. The penalized MLE in (11) also uses the notation p. We now replace it by p_0, to differentiate from the true sentiment p. In other words, we write

$$\widehat{p} = \arg\max_{p_0 \in [0,1]} \left\{ s^{-1} \sum_{j \in S} d_j \log\left(p_0 \widehat{O}_{+,j} + (1-p_0)\widehat{O}_{-,j}\right) + \lambda \log\left(p_0(1-p_0)\right) \right\}.$$

Let $\eta_0 = 2p_0 - 1$. It is seen that $p_0(1-p_0) = (1-\eta_0^2)/4$ and $p_0\widehat{O}_{+,j} + (1-p_0)\widehat{O}_{-,j} = \widehat{F}_j + \eta_0 \widehat{T}_j$. With re-parametrization by η_0, we have $\widehat{p} = (1+\widehat{\eta})/2$, where $\widehat{\eta} = \mathrm{argmax}_{\eta_0 \in [-1,1]} \widehat{\ell}(\eta_0)$, with

$$\widehat{\ell}(\eta_0) \equiv s^{-1} \sum_{j \in S} d_j \log(\widehat{F}_j + \eta_0 \widehat{T}_j) + \lambda \log(1-\eta_0) + \lambda \log(1+\eta_0). \tag{B.34}$$

Let $\eta^* = 2p^* - 1$. Then, $|\widehat{p} - p^*| \le |\widehat{\eta} - \eta^*|/2$. It is sufficient to bound $|\widehat{\eta} - \eta^*|$.

In preparation, we study $\widehat{\ell}'(\eta_0)$ and $\widehat{\ell}''(\eta_0)$. By direct calculations,

$$\widehat{\ell}'(\eta_0) = \sum_{j \in S} \frac{s^{-1} d_j \widehat{T}_j}{\widehat{F}_j + \eta_0 \widehat{T}_j} - \frac{2\lambda\eta_0}{1-\eta_0^2}, \qquad \widehat{\ell}''(\eta_0) = -\sum_{j \in S} \frac{s^{-1} d_j \widehat{T}_j^2}{(\widehat{F}_j + \eta_0 \widehat{T}_j)^2} - \frac{2\lambda(1+\eta_0^2)}{(1-\eta_0^2)^2}. \tag{B.35}$$

First, consider $\widehat{\ell}'(\eta_0)$. By Model (3), $d \sim \mathrm{Multinomial}(s, F + \eta T)$. It implies that $\mathbb{E}d_j = s(F_j + \eta T_j)$. Additionally, by Theorem 2, $(\widehat{F}_j, \widehat{T}_j)$ are close to $(F_j, \widehat{\rho} T_j)$. In light of this, we write

$$\widehat{f}(\eta_0) = \sum_{j \in S} \frac{s^{-1} d_j \widehat{T}_j}{\widehat{F}_j + \eta_0 \widehat{T}_j}, \qquad f(\eta_0) = \sum_{j \in S} \frac{(F_j + \eta T_j)\widehat{\rho} T_j}{F_j + \eta_0 \widehat{\rho} T_j},$$

where $\widehat{f}(\eta_0)$ is the first term in $\widehat{\ell}'(\eta_0)$ and $f(\eta_0)$ is its counterpart. Let E be the event that

$$|\widehat{F}_j - F_j| \le C\sqrt{F_j \log(m)/(n\bar{s})}, \qquad |\widehat{T}_j - \widehat{\rho} T_j| \le C\sqrt{F_j \log(m)/(n\bar{s})}, \tag{B.36}$$

simultaneously for all $j \in S$. In the proof of Theorem 2, we have seen that $\mathbb{P}(E) = 1 - o(1)$. We now condition on training data and assume that the realized $(\widehat{F}, \widehat{T})$ satisfy (B.36). Note that

$$\begin{aligned}
|\widehat{f}(\eta_0) - f(\eta_0)| &= \left| \sum_{j\in S} \frac{s^{-1} d_j \widehat{T}_j}{\widehat{F}_j + \eta_0 \widehat{T}_j} - \sum_{j\in S} \frac{s^{-1}(\mathbb{E}d_j)\,\widehat{\rho} T_j}{F_j + \eta_0 \widehat{\rho} T_j} \right| \\
&\le \left| \sum_{j\in S} \frac{s^{-1}(d_j - \mathbb{E}d_j)\widehat{T}_j}{\widehat{F}_j + \eta_0 \widehat{T}_j} \right| + \left| \sum_{j\in S} \frac{s^{-1}(\mathbb{E}d_j)(\widehat{T}_j - \widehat{\rho} T_j)}{\widehat{F}_j + \eta_0 \widehat{T}_j} \right| \\
&\quad + \left| \sum_{j\in S} s^{-1}(\mathbb{E}d_j)\widehat{\rho} T_j \left(\frac{1}{\widehat{F}_j + \eta_0 \widehat{T}_j} - \frac{1}{F_j + \eta_0 \widehat{\rho} T_j} \right) \right|.
\end{aligned}$$

First, note that $|\widehat{\rho}| \le 1$, $|\eta_0| \le 1$ and $|T_j| \le (1-c_1)F_j$. As a result, $F_j + \eta_0 \widehat{\rho} T_j \ge F_j - (1-c_1)F_j \ge c_1 F_j$. Second, by (B.36) and the assumption $n\bar{s}F_j/\log(m) \to \infty$, we know that $\max\{|\widehat{F}_j - F_j|, |\widehat{T}_j - \widehat{\rho} T_j|\} = o(F_j)$. It implies $\widehat{F}_j + \eta_0 \widehat{T}_j \ge (F_j + \eta_0 \widehat{\rho} T_j)[1 + o(1)] \gtrsim c_1 F_j$. Last, $s^{-1}\mathbb{E}d_j = F_j + \eta T_j \le 2F_j$. Plugging them into the above equation gives

$$\begin{aligned}
|\widehat{f}(\eta_0) - f(\eta_0)| &\le \left| \sum_{j\in S} \frac{s^{-1}(d_j - \mathbb{E}d_j)\widehat{T}_j}{\widehat{F}_j + \eta_0 \widehat{T}_j} \right| + C \sum_{j\in S} \sqrt{\frac{F_j \log(m)}{n\bar{s}}}, \\
&\le \left| \sum_{j\in S} \frac{s^{-1}(d_j - \mathbb{E}d_j)\widehat{T}_j}{\widehat{F}_j + \eta_0 \widehat{T}_j} \right| + C \sqrt{\frac{|S| \log(m)}{n\bar{s}}},
\end{aligned} \tag{B.37}$$

where the last inequality is due to $\sum_{j\in S} \sqrt{F_j} \le \sqrt{|S| \sum_j F_j} \le \sqrt{|S|}$. We then bound the first term in (B.37). We condition on the training data. Since d is independent of training data, it is still true that $d \sim \text{Multinomial}(s, F + \eta T)$. Let $\{b_\ell\}_{\ell=1}^s$ be iid random vectors, where $b_\ell \sim \text{Multinomial}(1, F + \eta T)$. Then, conditioning on the training data, d has the same distribution as $\sum_{\ell=1}^s b_\ell$. It follows that

$$\sum_{j\in S} \frac{s^{-1}(d_j - \mathbb{E}d_j)\widehat{T}_j}{\widehat{F}_j + \eta_0 \widehat{T}_j} \overset{(d)}{=} \sum_{\ell=1}^s \xi_\ell, \qquad \text{with} \quad \xi_\ell \equiv \sum_{j\in S} \frac{s^{-1}\widehat{T}_j}{\widehat{F}_j + \eta_0 \widehat{T}_j}(b_{\ell,j} - \mathbb{E}b_{\ell,j}).$$

Note that $(\widehat{F}, \widehat{T}, \widehat{\rho})$ are non-random, conditional on the training data. Hence, $\{\xi_\ell\}_{\ell=1}^s$ are iid variables with mean zero. We compute the variance of ξ_ℓ. Let $v \in \mathbb{R}^{|S|}$ be the vector such that $v_j = \frac{s^{-1}\widehat{T}_j}{\widehat{F}_j + \eta_0 \widehat{T}_j}$. Then, $\xi_\ell = v'(b_\ell - \mathbb{E}b_\ell)$. By elementary properties of multinomial distributions, the covariance matrix of ξ_ℓ is $\text{diag}(F + \eta T) - (F + \eta T)(F + \eta T)' \preceq \text{diag}(F + \eta T)$. Hence, $\text{Var}(\xi_\ell) \le \sum_{j\in S} v_j^2 (F_j + \eta T_j)$. It follows that

$$\text{Var}(\xi_\ell) \le \sum_{j\in S} \frac{s^{-2}\widehat{T}_j^2 \cdot (F_j + \eta T_j)}{(\widehat{F}_j + \eta_0 \widehat{T}_j)^2} \le C s^{-2} \sum_{j\in S} \frac{\widehat{\rho}^2 T_j^2}{F_j} \le C s^{-2} \widehat{\rho}^2 \Theta,$$

where we have used $|\widehat{T}_j| \le \widehat{\rho}|T_j| + o(F_j)$ and $\widehat{F}_j + \eta_0 \widehat{T}_j \gtrsim c_1 F_j$ (derivation is in the paragraph above

(B.37)). Additionally, $|\xi_\ell| \le \widehat{\rho}(sc_1)^{-1}\sum_{j\in S}|b_{\ell,j}| \le \widehat{\rho}(sc_1)^{-1}$. We apply (B.6) and find that, for any fixed η_0, conditioning on the training data, with probability $1-\epsilon$,

$$\Big|\sum_{j\in S}\frac{s^{-1}(d_j-\mathbb{E}d_j)\widehat{T}_j}{\widehat{F}_j+\eta_0\widehat{T}_j}\Big| \le C\Big(\sqrt{\frac{\widehat{\rho}^2\Theta\log(\epsilon^{-1})}{s}}+\frac{\widehat{\rho}\log(\epsilon^{-1})}{s}\Big) \le C\widehat{\rho}\sqrt{\frac{\Theta\log(\epsilon^{-1})}{s}}, \tag{B.38}$$

where the last inequality is due to $s\Theta\to\infty$. A combination of (B.37) and (B.38) gives the following result: Conditioning on the training data where the event E occurs (this event, as defined in (B.36), is about the training data; hence, it does not affect the probability here), for any given $\eta_0\in[-1,1]$, with probability $1-\epsilon$,

$$|\widehat{f}(\eta_0)-f(\eta_0)| \le C\Big(\widehat{\rho}\sqrt{\frac{\Theta\log(\epsilon^{-1})}{s}}+\sqrt{\frac{|S|\log(m)}{n\bar{s}}}\Big), \tag{B.39}$$

where the constant C does not depend on η_0. We now investigate $f(\eta_0)$. By definition, $\sum_{j\in S}F_j=1$ and $\sum_{j\in S}T_j=0$. Additionally, we recall that $\eta^*=\widehat{\rho}^{-1}\eta$. It follows that

$$\begin{aligned} f(\eta_0) &= \sum_{j\in S}\frac{(F_j+\eta_0\widehat{\rho}T_j)\widehat{\rho}T_j}{F_j+\eta_0\widehat{\rho}T_j}+\sum_{j\in S}\frac{(\widehat{\rho}^{-1}\eta-\eta_0)\widehat{\rho}^2T_j^2}{F_j+\eta_0\widehat{\rho}T_j} \\ &= \widehat{\rho}\sum_{j\in S}T_j+(\eta^*-\eta_0)\sum_{j\in S}\frac{\widehat{\rho}^2T_j^2}{F_j+\eta_0\widehat{\rho}T_j} \\ &= (\eta^*-\eta_0)\sum_{j\in S}\frac{\widehat{\rho}^2T_j^2}{F_j+\eta_0\widehat{\rho}T_j}. \end{aligned} \tag{B.40}$$

We plug (B.39) and (B.40) into $\widehat{\ell}'(\eta_0)$ in (B.35). It implies that, conditioning on training data where the event E occurs, for a fixed $\eta_0\in(-1,1)$, with probability $1-\epsilon$,

$$\Big|\widehat{\ell}'(\eta_0)-\Big[(\eta^*-\eta_0)\sum_{j\in S}\frac{\widehat{\rho}^2T_j^2}{F_j+\eta_0\widehat{\rho}T_j}-\frac{2\lambda\eta_0}{1-\eta_0^2}\Big]\Big| \le C\Big(\widehat{\rho}\sqrt{\frac{\Theta\log(\epsilon^{-1})}{s}}+\sqrt{\frac{|S|\log(m)}{n\bar{s}}}\Big). \tag{B.41}$$

Second, consider $\widehat{\ell}''(\eta_0)$. Since $d_j\ge 0$ and $\widehat{F}_j+\eta_0\widehat{T}_j\le 2F_j$ for all $\eta_0\in[-1,1]$, the first term of $\widehat{\ell}''(\eta_0)$ in (B.35) satisfies that

$$\sup_{\eta_0\in[-1,1]}\Big\{-\sum_{j\in S}\frac{s^{-1}d_j\widehat{T}_j^2}{(\widehat{F}_j+\eta_0\widehat{T}_j)^2}\Big\} \le -\sum_{j\in S}\frac{s^{-1}d_j\widehat{T}_j^2}{4\widehat{F}_j^2}.$$

Similar to (B.38), we can easily prove that, conditioning on the training data, with probability $1-\epsilon$,

$$\Big|\sum_{j\in S}\frac{s^{-1}(d_j-\mathbb{E}d_j)\widehat{T}_j^2}{4\widehat{F}_j^2}\Big| \le C\widehat{\rho}^2\sqrt{\frac{\Theta\log(\epsilon^{-1})}{s}}.$$

Combing the above and noting that $s^{-1}\mathbb{E}d_j = F_j + \eta T_j \geq c_1 F_j$, we have

$$\sup_{\eta_0 \in [-1,1]} \left\{ -\sum_{j \in S} \frac{s^{-1} d_j \widehat{T}_j^2}{(\widehat{F}_j + \eta_0 \widehat{T}_j)^2} \right\} \leq -\sum_{j \in S} \frac{c_1 F_j \widehat{T}_j^2}{4\widehat{F}_j^2} + C\widehat{\rho}^2 \sqrt{\frac{\Theta \log(\epsilon^{-1})}{s}}. \tag{B.42}$$

On the event E, $|\widehat{T}_j^2 - \widehat{\rho}^2 T_j^2| \leq C\widehat{\rho}|T_j||\widehat{T}_j - \widehat{\rho} T_j| \leq C|T_j|\sqrt{F_j \log(m)/(n\bar{s})}$, and similarly, $|\widehat{F}_j^2 - F_j^2| \leq CF_j\sqrt{F_j \log(m)/(n\bar{s})}$. It follows that,

$$\begin{aligned}
-\sum_{j \in S} \frac{c_1 F_j \widehat{T}_j^2}{4\widehat{F}_j^2} &\leq -\sum_{j \in S} \frac{c_1 F_j \widehat{\rho}^2 T_j^2}{4F_j^2} + \sum_{j \in S} c_1 F_j \widehat{\rho}^2 T_j^2 \Big| \frac{1}{4F_j^2} - \frac{1}{4\widehat{F}_j^2} \Big| + \sum_{j \in S} \frac{c_1 F_j |\widehat{T}_j^2 - \widehat{\rho}^2 T_j^2|}{4\widehat{F}_j^2} \\
&\leq -\sum_{j \in S} \frac{c_1 \widehat{\rho}^2 T_j^2}{4F_j} + C \sum_{j \in S} \sqrt{\frac{F_j \log(m)}{n\bar{s}}} \\
&\leq -\frac{c_1 \widehat{\rho}^2}{4} \Theta + C\sqrt{\frac{|S| \log(m)}{n\bar{s}}},
\end{aligned}$$

where the last line is due to $\sum_{j \in S} \sqrt{F_j} \leq \sqrt{|S| \sum_{j \in S} F_j} \leq \sqrt{|S|}$. We plug it into (B.42) to get

$$\sup_{\eta_0 \in (-1,1)} \left\{ -\sum_{j \in S} \frac{s^{-1} d_j \widehat{T}_j^2}{(\widehat{F}_j + \eta_0 \widehat{T}_j)^2} \right\} \leq -\frac{c_1 \widehat{\rho}^2}{4} \Theta + C\left(\widehat{\rho}^2 \sqrt{\frac{\Theta \log(\epsilon^{-1})}{s}} + \sqrt{\frac{|S| \log(m)}{n\bar{s}}} \right) \tag{B.43}$$

We further plug it into the expression of $\widehat{\ell}''(\eta_0)$ in (B.35). It implies that, conditioning on the training data where the event E occurs, with probability $1 - \epsilon$, simultaneously for all $\eta_0 \in (-1, 1)$,

$$\widehat{\ell}''(\eta_0) \leq -\frac{c_1 \widehat{\rho}^2}{4} \Theta - \frac{2\lambda(1 + \eta_0^2)}{(1 - \eta_0^2)^2} + C\left(\widehat{\rho}^2 \sqrt{\frac{\Theta \log(\epsilon^{-1})}{s}} + \sqrt{\frac{|S| \log(m)}{n\bar{s}}} \right). \tag{B.44}$$

With (B.41) and (B.44), we are ready to show the claim. Since $\lambda > 0$, it is true that $\widehat{\ell}(\eta_0) \to -\infty$ as $\eta_0 \to \pm 1$. Therefore, the maximum can only be attained in the interior of $(-1, 1)$. By Lagrange's mean value theorem, there exists $\widetilde{\eta}$, which is between η^* and $\widehat{\eta}$, such that

$$0 = \widehat{\ell}'(\widehat{\eta}) = \widehat{\ell}'(\eta^*) + \widehat{\ell}''(\widetilde{\eta}) \cdot (\widehat{\eta} - \eta^*).$$

It follows that

$$|\widehat{\eta} - \eta^*| \leq \frac{|\widehat{\ell}'(\eta^*)|}{|\widehat{\ell}''(\widetilde{\eta})|} \leq \frac{|\widehat{\ell}'(\eta^*)|}{\left(\inf_{\eta_0 \in (-1,1)} |\widehat{\ell}''(\eta_0)| \right)}. \tag{B.45}$$

Recall that $err_n = \Theta^{-1}\left(\sqrt{|S| \log(m)/(n\bar{s})} + \sqrt{\Theta/s} \right)$ and note that $0 < \widehat{\rho} \leq 1$. Then, in both (B.41) and (B.44), the term in the brackets can be written as $C_\epsilon \Theta err_n$, where C_ϵ is a generic constant that depends on ϵ, the meaning of which varies from occurrence to occurrence. Additionally, $|\eta^*| \leq 1 - 2c_2$,

by our assumption. It follows from (B.41) and (B.44) that

$$|\widehat{\ell}(\eta^*)| \le \frac{2\lambda|\eta^*|}{1-(\eta^*)^2} + C_\epsilon \Theta err_n \le \frac{\lambda|\eta^*|}{c_2} + C_\epsilon \Theta err_n,$$
$$\inf_{\eta_0\in(-1,1)} |\widehat{\ell}'(\eta_0)| \ge \inf_{\eta_0\in(-1,1)} \Big\{ \frac{c_1\widehat{\rho}^2\Theta}{4} + \frac{2\lambda(1+\eta_0^2)}{(1-\eta_0^2)^2} \Big\} - C_\epsilon\Theta err_n \ge \frac{c_1\rho_0^2\Theta}{4} + 2\lambda - C_\epsilon\Theta err_n.$$

According to (B.41) and (B.44), the above inequalities hold with probability $1-\epsilon$ when conditioning on the training data where the event E occurs. Since the right hand side does not depend on training data, the above holds with the same probability marginally (i.e., without conditioning), provided that the event E occurs. We have seen that $\mathbb{P}(E) = 1 - o(1)$. Therefore, the above inequalities hold with probability $1-\epsilon$. We plug them into (B.45) and find that, with probability $1-\epsilon$,

$$\begin{aligned}
|\widehat{\eta} - \eta^*| &\le \frac{4\lambda|\eta^*| + C_\epsilon\Theta\cdot err_n}{8c_2\lambda + c_1c_2\rho_0^2\Theta - C_\epsilon\Theta err_n} \\
&\le C_\epsilon \begin{cases} \Theta^{-1}\lambda|\eta^*| + err_n, & \text{if } \lambda \le \Theta, \\ |\eta^*| + \lambda^{-1}\Theta err_n, & \text{if } \lambda > \Theta, \end{cases} \\
&\le C_\epsilon \min\Big\{1, \frac{\lambda}{\Theta}\Big\} \times |\eta^*| + C_\epsilon \min\Big\{1, \frac{\Theta}{\lambda}\Big\} \times err_n. \qquad \text{(B.46)}
\end{aligned}$$

This proves the claim. □

B.6 Proof of Theorem 4

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

$$SR(\hat{p}, p) = 1 - \frac{6}{L(L^2-1)} \sum_{i=1}^{L} (\hat{r}_i - r_i)^2, \qquad \text{(B.47)}$$

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

$$r_i = \frac{1}{2}\sum_{j=1}^{L} \mathrm{sgn}(p_i^* - p_j^*) + \frac{L+1}{2}, \qquad \hat{r}_i = \frac{1}{2}\sum_{j=1}^{L} \mathrm{sgn}(\hat{p}_i - \hat{p}_j) + \frac{L+1}{2},$$

where the sign function takes values in $\{0, \pm 1\}$. We condition on $\{p_i\}_{i=1}^L$ and let $\epsilon = L^{-2}$ in Theorem 3 (the statement of this theorem assumes that ϵ is a fixed constant, but the proof can be easily extended to allow for $\epsilon \to 0$) and apply the probability union bound. It yields that, with probability $1 - L^{-1}$,

$$\max_{1\le i\le L} |\hat{p}_i - p_i^*| \le \delta, \qquad \text{where} \quad \delta = \frac{C}{\sqrt{\Theta}} \Big(\frac{\sqrt{|S|\log(m)}}{\sqrt{n\bar{s}\Theta}} + \frac{\sqrt{\log(L)}}{\sqrt{s}} \Big). \qquad \text{(B.48)}$$

Let D be the event that (B.48) holds. For each $1 \le i \le L$, define the index set

$$B_i(3\delta) = \{1 \le j \le L : j \ne i,\ |p_j^* - p_i^*| \le 3\delta\}.$$

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

$$|\hat{r}_i - r_i| \le \frac{1}{2} \sum_{j \in B_i(3\delta)} \big(|\mathrm{sgn}(p_i^* - p_j^*)| + |\mathrm{sgn}(\hat{p}_i - \hat{p}_j)|\big) \le |B_i(3\delta)|.$$

We plug it into (B.47) and note that $|\hat{r}_i - r_i|^2 \le L|\hat{r}_i - r_i|$. It yields

$$1 - SR(\hat{p}, p) \le \frac{6}{L^2 - 1} \sum_{i=1}^{L} |\hat{r}_i - r_i| \le \frac{6L}{L^2 - 1} \max_{1 \le i \le L} |B_i(3\delta)|. \tag{B.49}$$

More precisely, conditioning on $\{p_i\}_{i=1}^L$, (B.49) holds with probability $1 - L^{-1}$.

We now bound $|B_i(3\delta)|$, taking into account the randomness of $\{p_i\}_{i=1}^L$. Each p_i^* is a non-random linear function of p_i. Therefore, the distribution of $\{p_i\}_{i=1}^L$ yields that $\{p_i^*\}_{i=1}^L$ are *iid* drawn from a continuous density on $[\frac{1}{2} - (\rho_0/\widehat{\rho})(\frac{1}{2} - c_2),\ \frac{1}{2} + (\rho_0/\widehat{\rho})(\frac{1}{2} - c_2)]$ (note: $\widehat{\rho}$ depends on the training data, hence, it is independent of p_i^*'s). Since this is a compact set, the probability density function must be Lipschitz. Fix i and write

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

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 $(L-1)$ *iid* Bernoulli variables with a success probability of $O(\delta)$. Then, $\mathbb{E}|B_i(3\delta)| = O(L\delta)$. Moreover, by the Bernstein's inequality (B.6), with probability $1 - L^{-2}$,

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

Combining it with the probability union bound, with probability $1 - L^{-1}$, the above inequality holds simultaneously for all $1 \le i \le L$. We then plug it into (B.49) and get

$$1 - SR(\hat{p}, p) \le C\delta + C\sqrt{\frac{\delta \log(L)}{L}} + \frac{C\log(L)}{L} \le C \max\Big\{\delta,\ \frac{\log(L)}{L}\Big\}. \tag{B.50}$$

Under our assumption, the right hand side of (B.50) is $o(1)$. The claim follows immediately. □

C 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 follows:

$$d_{[S],i} \sim \text{Multinomial}\Big(s_i,\ p_i O_+ + (1-p_i)O_-\Big), \quad d_{[N],i} \sim \text{Multinomial}\Big(n_i,\ \Omega\Big), \tag{C.51}$$

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 Ω_j is drawn from $\frac{1}{m-|S|}\text{Unif}(0,2)$, for $j = |S|+1, \ldots, m$, then renormalized such that $\sum_j \Omega_j = 1$. 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. As such, the difference between $O_{+,j}$ and $O_{-,j}$ reaches the minimum at $j = |S|/2$. This means among all words within the set S, those with j around $|S|/2$ are close to neutral.

Next, the sign of returns follows a logistic regression model: $\mathbb{P}(y_i > 0) = p_i$, and its magnitude $|y_i|$ follows a folded Student t-distribution with the degree of freedom parameter set at 4. The standard deviation of the t-distribution has negligible effects on our simulations, since only the ranks of returns matter for our results.

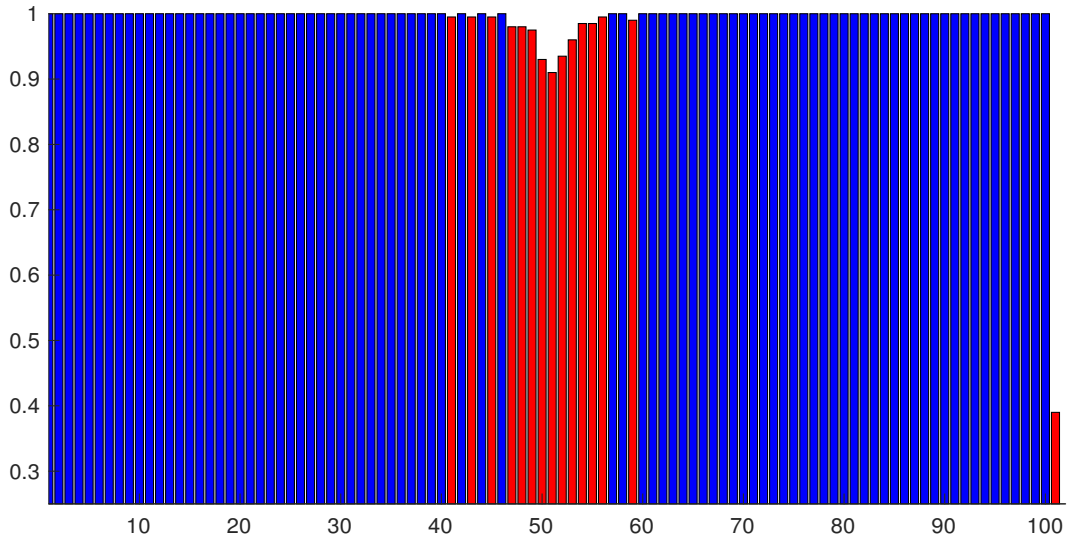
We fix the number of Monte Carlo repetitions at $M_c = 200$ and the number of articles in the testing sample at $1,000$. In the benchmark case, we set $|S| = 100$, $m = 500$, $n = 10,000$, $\bar{s} = 10$, and $\bar{n} = 100$.

We first evaluate the screening step. Although tuning threshold parameters give better results, for convenience we choose the top 100 words in terms of $|f_j - 0.5|1_{\{k_j > \kappa\}}$, where κ is set at the 10% quantiles of all k_js. Figure A.1 reports the frequency of each word selected in the screening step across all Monte Carlo repetitions. Across all repetitions, the probability of selecting any word outside the set S is about 0.4%. Not surprisingly, the words in S that are occasionally missed are those with corresponding js around $|S|/2$.

Next, Figure A.2 illustrates the accuracy of the estimation step, taking into account potential errors in the screening step. The true values of T and F are shown in black. The scaling constant ρ is approximately 0.5 in this 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 for those words in F that are occasionally missed from the screening step.

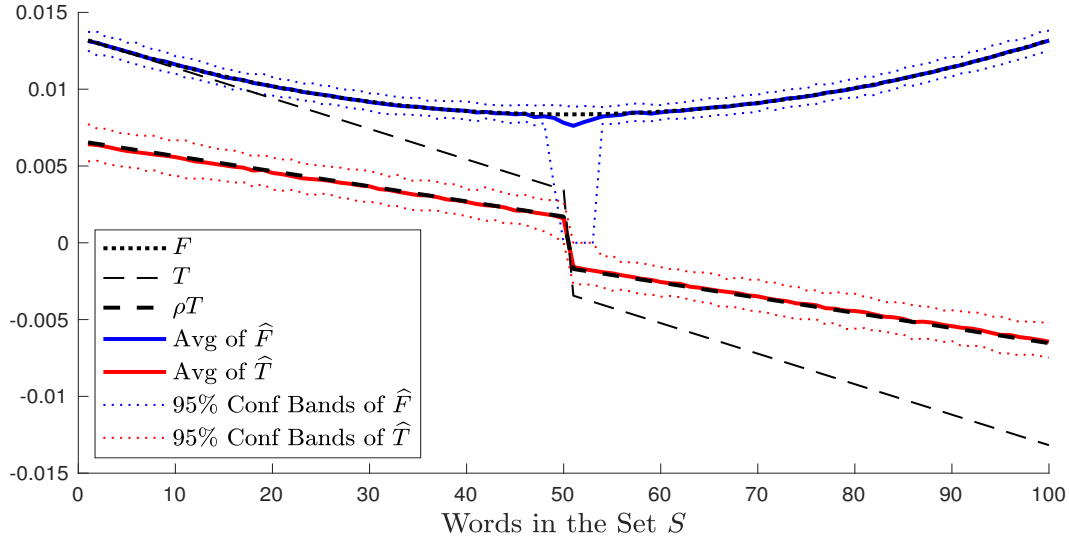
Finally, we examine the accuracy of the scoring step, accounting for errors propagated from the previous steps. Data from the testing sample are never used in the previous two steps. Table A.1 reports Spearman's rank correlation coefficients between the predicted $\widehat{p}$ and the true p for 1,000 articles in the testing sample in a variety of cases. We report the rank correlation because what

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.

matters is the rank of all articles instead of their actual scores (which are difficult to consistently estimate due to the bias of estimating $\widehat{p}_i$). Also, the penalization term ($\lambda = 0.5$) in our likelihood shrinks the estimated scores towards 0.5, although it barely has any 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 having fewer observations from words in S. Similarly, if we decrease n to 5,000 the estimates become less accurate because the sample size is smaller. If the size of the vocabulary, m, or the size of the dictionary of the sentiment words, $|S|$, drop by half, the estimates improve, though the improvement is marginal. Overall, these observations match what our theory predicts.

Table A.1: Spearman's Correlation Estimates

	benchmark	$\bar{s} \downarrow$	$n \downarrow$	$m \downarrow$	$\lvert S\rvert \downarrow$
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

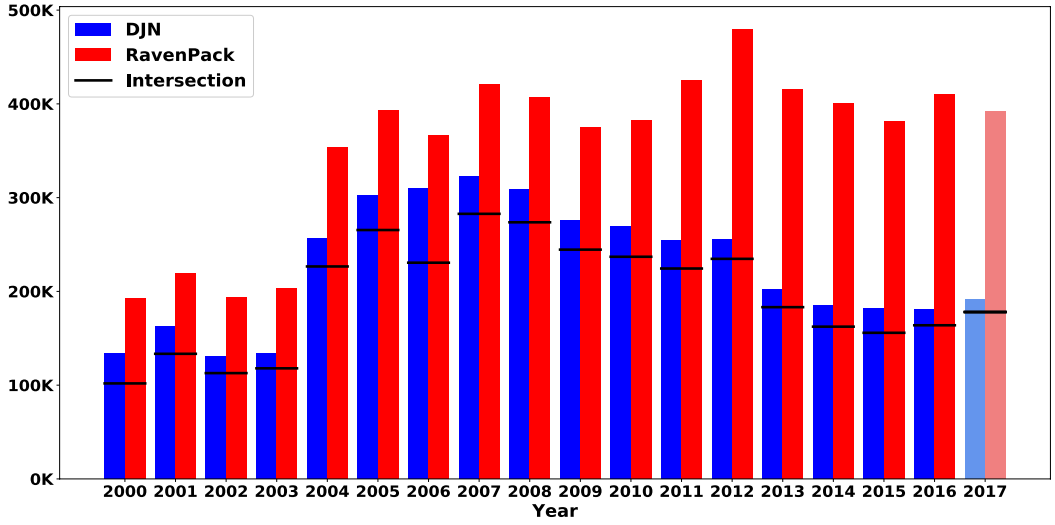
Note: In this table, we report the mean and standard deviation of Spearman's correlation estimates across Monte Carlo repetitions for a variety of cases. The parameters in the benchmark case are set as: $|S| = 100$, $m = 500$, $n = 10,000$, and $\bar{s} = 10$. In each of the remaining columns, the corresponding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

D 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 SESTM. 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 its analytics process. Ultimately, what we observe in RavenPack is their collection of article-level scores 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.

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 counts 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.

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



E 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.