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

December 16, 2018

Abstract

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

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 gratefully acknowledge the computing support from the Research Computing Center at the University of Chicago. AQR Capital Management is a global investment management firm, which may or may not apply similar investment techniques or methods of analysis as described herein.

1 Introduction

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

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

Highly influential studies in this vein include Tetlock (2007) and Loughran and MacDonald (2011). These papers manage the dimensionality challenge of text by avoiding the need to estimate a model for the statistical behavior of the text itself. Using word scores from pre-existing sentiment dictionaries is equivalent to using model estimates from a past study to construct fitted values in a new collection of documents being analyzed. This approach has the great advantage that it allows researchers to make progress on understanding certain aspects of the data without taking on the (often onerous) task of estimating a model for a new text corpus from scratch.

The goal of this paper is to demonstrate how basic machine learning techniques can be used to understand the sentimental structure of a text corpus without relying on pre-existing dictionaries. The method we suggest has three main virtues. The first is simplicity—it requires only standard econometric techniques such as calculating correlations and maximum likelihood estimation. Second, our method requires minimal computing power—it can be run with a laptop computer in a matter of minutes for text corpora with millions of documents. Third, and most importantly, it allows the researcher to construct a sentiment scoring model that is specifically adapted to the context of the dataset at hand. This frees the researcher from relying on a pre-existing sentiment dictionary that was originally designed for different purposes.

Our empirical analysis revisits perhaps the most commonly studied text-based research questions 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*, one of the most widely subscribed and closely monitored news sources by market participants. It contains time-stamped news articles over a 38-year time span. The dataset tags articles with the identifiers of firms to which the article pertains. We merge this with CRSP stock return data, which allows us to

model stock-level return behavior as a function of a newswire article's content. From this data, we learn the sentiment scoring model. A key aspect of our approach is that we estimate the model from the joint behavior of newswire text and stock returns. Covariation among prices and word usage reveal the sentiment structure of the text as it applies to the specific context of return forecasting. Our estimated model isolates an easily interpretable and highly intuitive list of positive and negative sentiment words and their associated sentiment scores. We demonstrate the predictive capacity of our 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).

1.1 Methodological Contributions

We call our procedure SSESTM (pronounced "system," for Supervised Sentiment Extraction via Screening and Topic Modeling). The model consists of three parts, and basic machine learning strategies play a central role in each.

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

The second step is to assign term-specific sentiment weights based on their individual relevance for the prediction task. Text data is typically well approximated by Zipf's law, which predicts a small number of very high-frequency terms and a very large number of low-frequency terms. While existing finance literature recognizes the importance of accounting for vast differences in term frequencies when assigning sentiment weights, the ultimate choice of weights has typically been ad hoc (e.g., via term frequency-inverse document frequency weighting, as used by Loughran and McDonald (2011)). We instead use a likelihood-based, or "generative," model to account for the extreme skewness in term frequencies. The specific machine learning tool we apply in this component is a supervised topic model. For the sake of simplicity and computational ease, and because it is well adapted to our purposes, we opt for simple two-topic model—one that describes the frequency distribution of positive sentiment terms, and one for negative sentiment words.

Finally, we use the estimated topic model to assign an article-level polarity score. When aggregating to an article score, we use the internally consistent likelihood structure of the model to account for the severe heterogeneity in both the frequency of words as well as their sentiment weights when aggregating. In light of the low signal-to-noise ratio in asset return data, we design a simple

penalized maximum likelihood estimator, with only one unknown parameter to estimate for each article. A Bayesian interpretation of the penalization is to impose a Beta-distributed prior on the polarity score that centers at 1/2, i.e., a neutral sentiment, because most of the articles have a neutral tone.

We establish the theoretical underpinnings of all these algorithms, and in particular shed light on their bias, if any, and statistical efficiency, and how they depend on the length of the dictionary, the number of news articles, and their average length.

Our framework only uses the word count of each article, ignoring information about the sequence of words. This so-called bag of words approach to text mining is probably not most effective, because the sequence of words contains information about the context in which the words are used, and that the context of words typically matters for their meaning. We view this as a compromise for simplicity, transparency, and interpretability. We leave applications of more sophisticated approaches that incorporate information about context, such as the deep learning approach, to future work.

1.2 Literature

There is a burgeoning strand of literature that apply machine learning techniques to asset pricing, see, e.g., Freyberger et al. (2017), Kozak et al. (2017), Kelly et al. (2017), Feng et al. (2017). In particular, Gu et al. (2018) review a suite of machine learning tools in the hope of exhausting the return predictability buried in the fundamental data of individual firms. Their conservative estimates of the economic gains point out that machine learning holds promises in predicting both monthly and annual returns for the cross-section of individual firms and aggregated portfolios, respectively. In contrast, our paper focuses on the use of alternative data, the news text, as well as the prediction of daily returns.

The existing literature on text sentiment analysis in accounting and finance largely use pre-existing or self-created dictionaries, see a review article by Loughran and Mcdonald (2016). Although these studies have shown that certain phrases have statistically significant predictive power for future returns, the economic magnitude is often shown to be limited. Using Harvard IV-4 psychosocial dictionary, Tetlock (2007) conducts an in-sample investigation of Wall Street Journal columns and finds that a one standard deviation increase in pessimism is related to only an 8.1 basis point decline in the Dow the following day. Relying on the same dictionary, Tetlock et al. (2008) further predict individual S&P 500 firms' accounting earnings and returns using relative frequency of negative words in each news story, and find the fraction of negative words in firm-specific news stories generally have very little ability to forecast future returns. Loughran and McDonald (2011) use a self-created dictionary based on words used in firms' 10-K filings, and find stronger association between their negative word lists and firms' returns at the time of a 10-K filing. They do not however, explore the predictive performance of the association. We construct and evaluate the performance of a trading strategy that relies on our sentiment estimates. The economic gains is more substantial, thanks to the use of more powerful machine learning tools.

There are few exceptions in the literature that use supervised learning, such as the Naïve Bayes

method or support vector machines. Using Naïve Bayes approach, Antweiler and Frank (2005) find that internet stock messages posted on Yahoo Finance and Raging Bull about 45 companies help predict market volatility, but the effect on stock returns is statistically significant but economically small. Related work also include Li (2010), Jegadeesh and Wu (2013), and Huang et al. (2014). As pointed out in Loughran and Mcdonald (2016), Naïve Bayes approach based classification algorithms involve thousands of various unpublished rules/filters to measure the context of documents, and hence is less transparent and difficult to replicate. Our model is generative, transparent, and tractable. We also provide theoretical guarantees on all algorithms we propose.

Modern text mining in the computer science and machine learning literature has placed an increasing emphasis on unsupervised learning approaches, such as latent Dirichlet allocation (LDA) method for topic modeling by Blei et al. (2003) and its various extensions, vector representation of words and phrases, i.e., word2vec, by Mikolov et al. (2013), etc. We adopt a supervised approach thanks to the available information on the tagged firms associated with news articles and their returns. In the context of asset pricing, where the signal-to-noise ratio is notoriously low, it is often helpful to incorporate valuable information for supervision.

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

2 Methodology

2.1 Model Setup

We start with some notation. Suppose we have a collection of n news articles and a dictionary of m words. We record the word counts of the i-th article in a vector $d_i \in \mathbb{R}_+^m$, so that $d_{i,j}$ is the total number of word j in this article, for $1 \le i \le n, 1 \le j \le m$. In matrix form, we can write it as an $n \times m$ term-document matrix $D = (d_{i,j})$. We use the notation $a_{i,[S]}$ and $A_{\cdot,[S]}$ to denote a row vector corresponding to the ith row of A and a submatrix of A, respectively, whose columns are given by the set S.

We tag the stocks mentioned in the articles. For simplicity, we assume each article mentions only one stock, and associate this article with the stock return (or its idiosyncratic component) on the publication date of the article. In addition, we assume each article is given a polarity score $p_i \in [0, 1]$, only through which the article influences the stock return. That is,

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

where y_i is the stock return associated with the ith article.

There are certain limitations with this simple setup. First, it is possible that some articles might write about multiple names. For instance, Apple and Samsung are competitors in the cellphone market. There may exist certain news articles that draw a comparison between these two firms. The sentiment for them may be rather different. For simplicity, we eliminate articles with multiple stocks

mentioned in the empirical analysis. More complicated models are required to allow for multiple names in one article, which we leave for future work.

Second, we assume the publication date of the article is the same as the date on which the stock return is recorded. This is critical because we use the sign of the stock returns to label the sentiment of the articles. This assumption implies that the news must be "fresh" and the markets need to respond to the news. Also, the article cannot be retrospective. While this assumption is restrictive, our main empirical results use a three-day window around the news event to alleviate the concern.

In light of the conditional independence assumption, we need two additional components to fully specify the data generating process. The first component relates the stock return to the polarity of the article, that is, the distribution of y_i given p_i. The second component relates the word counts of the article to its polarity score, that is, the distribution of d_i given p_i.

We start with the first component. We assume

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

where sgn(x) is the sign function that returns 1 only if $x > 0$ and 0 otherwise. Intuitively, this assumption implies that the higher the polarity score, the higher the probability of realizing a positive return. Note that we do not need to specify the full distribution of y_i or the particular form of $g(\cdot)$.

We now turn to the second component. We assume the dictionary has a partition:

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

where S is the set of sentiment-sensitive words and N is the set of sentiment-neutral words. Let $d_{i,[S]}$ and $d_{i,[N]}$ be the corresponding subvector of d_i, and we assume they are generated independently of each other. To model $d_{i,[S]}$, we introduce two vectors of non-negative entries, O_+, $O_- \in \mathbb{R}_+^{|S|}$, with unit ℓ^1-norms, representing positive and negative opinions, respectively. They can be regarded as probability distributions on the space of sentiment-sensitive words. We thereby write the sentiment-sensitive word-count generating process as:

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

where s_i is the total counts of sentiment-sensitive words in the ith article. Equation (4) is essentially a 2-topic topic model. Our main interest is to learn the two distributions, O_+ and O_-, as well as the polarity score of each article p_i. The model of $d_{i,[N]}$ is largely a nuisance, so we leave as much flexibility as possible. We only assume that each $d_{i,[N]}$ follows a multinomial distribution, subject to a mild requirement that the counts of sentiment-neutral words and the article's polarity score cannot be too related.[1]

Our model is generative and semiparametric in essence. It fully specifies how sentiment-sensitive words are generated, yet is agnostic about the data generating process of the sentiment-neutral words.

[1] We may further model $d_{i,[N]}$ using a standard topic model (Hofmann, 1999; Blei et al., 2003), with K sentiment-irrelevant topics. This is, however, not necessary for the purpose of sentiment analysis.

In addition, the model only requires a monotonic assumption between the sentiment of the article and the sign of its tagged returns, yet is agnostic about the functional form or return distributions. In what follows, we develop machine learning procedures to estimate this model.

2.2 Identifying Sentiment-Sensitive Words

Because the sentiment-sensitive words are mixed with the sentiment-neutral words in the dictionary, directly estimating a topic model using the entire dictionary at best incurs a loss of efficiency, if not infeasible. The sentiment-neural words serve as noises in this model, and they likely dominate in quantity. We need an effective feature selection procedure to tease out words with sentiment.

We take a screening approach. We use stock returns to screen out sentiment-neutral words that serve as noises in the sentiment-topic model. Specifically, we outline below the procedure to build $\widehat{S}$, an estimate of S.

Algorithm 1.

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

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

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

$$\widehat{S} = \big\{j : f_j \geq 1/2 + \alpha_+\big\} \cup \big\{j : f_j \leq 1/2 - \alpha_-\big\} \cap \{j : k_j \geq \kappa\}, \tag{6}$$

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

S1 relies on the number of articles instead of the total number of word counts, so that a world is only counted once even if it is mentioned multiple times within one article. This is because in practice sentiment words are likely adjectives, which are less likely to appear multiple times in an article than nouns. In theory, both approaches are consistent.

In S2, in addition to the thresholds on the score f_j, we also impose a third threshold on the denominator of the f_j. This is because some sentiment words might happen to appear infrequently in the data sample we have, in which case we have very noisy information about their relevance to polarity, imposing this threshold is amount to taking into account the statistical accuracy of the score f_j.

We can also combine our vocabulary with words identified in the classical sentiment analysis, which has provided various lexical vocabularies for common English writing.[2]

[2]It is simple to expand $\widehat{S}$ to $\widetilde{S}$, by replacing S2 in Algorithm 1:

$$\widetilde{S} = \widehat{S} \cup \big\{1 \leq j \leq m : \max\{\ell_j, 1 - \ell_j\} \geq 1 - \beta\big\}. \tag{7}$$

where $\ell \in [0,1]^m$ is a vector capturing the lexical level of all dictionary words in common English writing, and $\beta > 0$ is a proper threshold we can tune.

To justify the validity of this algorithm, we establish the so-called sure-screening property (Fan and Lv, 2008) of Algorithm 1 in Theorem A.2 of Appendix A. In other words, we show $\mathbb{P}(\widehat{S} = S)$ approaches one asymptotically as n and m increase to ∞.

2.3 Learning Sentiment Distributions

We now have simplified the problem to a specific topic model with two topic vectors, summarized in a matrix $O = [O_+, O_-]$. We then construct its estimator $\widehat{O}$, which distributes on the space of $\widehat{S}$. The matrix O determines the data generating process of the word counts of each article, capturing information on the frequency of words as well as their sentiment. In fact, we rewrite the two sentiment vectors to a *vector of frequency* and a *vector of tone*:

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

where the vector of frequency characterizes the relative frequency of different words, whereas the vector of tone reflects the relative sentiment of different words. They are the key parameters that relate the sentiment of an article to the sentiment of each word therein. In this section, we provide a supervised learning approach to estimate O. Given any estimator of $\widehat{O}$, we can define $\widehat{F}$ and $\widehat{T}$ similarly based on (8).

Our setting differs from the classical topic models (Hofmann, 1999) in that each article is associated with a stock return. The returns can be used as training labels. Existing topic modeling methods, (Blei et al., 2003; Anandkumar et al., 2012; Arora et al., 2013; Ke and Wang, 2017), nonetheless, are unsupervised approaches, that ignore training labels. In a low signal-to-noise ratio environment, supervised learning often leads to superior performance.

Algorithm 2.

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

$$\widetilde{p}_i = \frac{\text{rank of } y_i \text{ in all returns}}{n}. \tag{9}$$

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

$$\widetilde{O} = \widetilde{D}\widetilde{W}'(\widetilde{W}\widetilde{W}')^{-1}, \qquad \text{where} \quad \widetilde{W} = \begin{bmatrix} \widetilde{p}_1 & \widetilde{p}_2 & \cdots & \widetilde{p}_n \\ 1-\widetilde{p}_1 & 1-\widetilde{p}_2 & \cdots & 1-\widetilde{p}_n \end{bmatrix}. \tag{10}$$

Set negative entries of $\widetilde{O}$ to zero and re-normalize each column to have a unit ℓ^1-norm. The resulting matrix is $\widehat{O}$.

We first explain the rationale of S2 in Algorithm 2. Note that model (4) implies $\mathbb{E}s_i^{-1}d_{i,[S]} =$

$p_i O_+ + (1-p_i)O_-$. In matrix form, we have

$$\mathbb{E}\widetilde{D} = OW, \qquad \text{where} \quad W = \begin{bmatrix} p_1 & \cdots & p_n \\ 1-p_1 & \cdots & 1-p_n \end{bmatrix}.$$

This suggests a recovery of O via regressing $\widetilde{D}$ on $\widetilde{W}$. S1 of Algorithm 2 uses the standardized ranks of returns as polarity scores. This approach is very intuitive, and robust to return outliers.

The accuracy of the estimator $\widehat{O}$ thereby hinges on the approximation quality of $\{\widetilde{p}_i\}_{i=1}^n$ with respect to $\{p_i\}_{i=1}^n$ as well as the quality of the wordlist $\widehat{S}$ obtained from screening. Theorem A.3 in Appendix A characterizes the statistical accuracy of the algorithm precisely.

It turns out the algorithm can consistently recover F, but with respect to T, a bias term may arise, which depends on the correlation between the true polarity and the estimated polarity:

$$\rho = \frac{12}{n}\sum_{i=1}^{n}\Big(p_i - \frac{1}{2}\Big)\Big(\widetilde{p}_i - \frac{1}{2}\Big). \tag{11}$$

Specifically, the estimator converges to ρT instead of T. That being said, as long as the estimation quality of $\widetilde{p}$ is high, ρ is fairly close to 1. It equals 1 if $\widetilde{p} = p$. Moreover, our trading strategies only depend on the relative rank of polarity scores of articles, instead of the absolute scores. The bias due to a scaling factor makes no difference in that regard.

Given n articles realized from our topic model, with a vocabulary of size $|S|$ (i.e., the number of words in S), and an average article length of $\bar{s}$, we show the convergence rate of the estimation errors of F and ρT are bounded by $\sqrt{|S|/(n\bar{s})}$, up to a logarithmic factor. In our empirical study, the identified sentiment dictionary contains $100 \sim 200$ words, yet their total count in one article is typically below 20. So we are primarily interested in the "short article" case, that is, $\bar{s}/|S| \le C$ for some constant C, as opposed to the "long article" case, in which $\bar{s}/|S| \to \infty$. A shown in Ke and Wang (2017), the classical unsupervised approach cannot match our rate in the case of short articles. This sheds light on the statistical efficiency gain of our supervised approach against unsupervised methods.

2.4 Scoring New Articles

In the previous steps, we have constructed estimators $\widehat{S}$ and $\widehat{O}$. We now discuss how to estimate the polarity p for a new article with word-count $d \in \mathbb{R}_+^m$, which is not involved in the previous estimation procedure.

By (4), we have

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

where s is the total number of words that appear in S in this article. To estimate p using $\widehat{S}$ and $\widehat{O}$, our approach is to adopt the maximum likelihood estimation (MLE). Although alternative estimators, such as linear regressions are also consistent, we use likelihood for its statistical efficiency, along with a penalty to regularize the estimate, given the limited amount of data. Because there is only one

parameter to estimate, a simple binary search is also computationally efficient.

Algorithm 3.

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

$$\widehat{p} = \arg\max_{p\in[0,1]} \left\{ \widehat{s}^{-1} \sum_{j=1}^{\widehat{s}} d_j \log\left(p\widehat{O}_{+,j} + (1-p)\widehat{O}_{-,j}\right) + \lambda \log\left(p(1-p)\right) \right\}, \tag{12}$$

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.

The first term on the right-hand side of (12) is the log likelihood function, barring a term that does not depend on p. The second term is a penalty which helps cope with the low signal-to-noise ratio environment. Imposing the penalty shrinks the estimate towards $1/2$, i.e., the neutral sentiment score. The amount of shrinkage depends on the magnitude of λ. This is also equivalent to imposing a prior Beta distribution on the polarity score. For most articles, their sentiment is neutral, so imposing such a prior improves the estimates. For articles that are not neutral, the polarity scores will be biased towards neutrality if λ is large (or the prior is not uninformative). Nonetheless, the relative ranks across articles are less likely influenced, which convey the most useful information for the trading strategy we build in the empirical analysis.

Theorem A.4 in Appendix A provides the statistical guarantee on the validity of the procedure. Perhaps not surprisingly, the estimator is inconsistent with respect to p, but converge to $\frac{1}{2}+\frac{1}{\rho}\left(p-\frac{1}{2}\right)$ instead. This $1/\rho$ inflating factor (ρ is often smaller than 1) arises because of the bias in the estimation of T. This is another reason why we impose penalization or shrinkage to help deflate the estimates. As we have discussed previously, this has no impact on the relative polarity scores of new articles, nor on our trading strategies. In terms of the convergence rate, besides the estimation error accumulated from the previous two steps, an additional error of magnitude $1/\sqrt{s}$ appears. This is intuitive because if the article contains very few sentiment words, its polarity score would not be accurately recovered.

2.5 Training, Validation, and Testing

Like most machine learning procedures, our algorithms rely on a choice of hyperparameters (or, synonymously, "tuning parameters"). There are four parameters $(\alpha_+, \alpha_-, \kappa, \lambda)$ we need to select.

We follow the most common approach in the literature and select tuning parameters adaptively from the data in a validation sample. In particular, we divide our sample into three disjoint time periods that maintain the temporal ordering of the data. The first, or "training," subsample is used to estimate the model subject to a specific set of tuning parameter values.

The second, or "validation," sample is used for tuning the hyperparameters. We construct polarity scores for news articles in the validation sample based on the estimated model from the training sample. Next, we search for hyperparameters that minimizes some loss function based on prediction errors from the validation sample. Our loss function is the sum of the absolute differences between predicted polarity scores and the (standardized) rank of the articles in the validation sample.

Tuning parameters are chosen from the validation sample taking into account estimated parameters, but the parameters are estimated from the training data alone. The idea of validation is to simulate an out-of-sample test of the model. Hyperparameter tuning amounts to searching for a degree of model complexity that tends to produce reliable out-of-sample performance. The validation sample fits are thereby not truly out-of-sample. Thus the third, or "testing," subsample, which is used for neither estimation nor tuning, is truly out-of-sample and thus is used to evaluate the method's predictive performance.

3 Empirical Analysis

3.1 Data and Preprocessing

We obtain machine text feed and archive database from the Dow Jones Newswires. The dataset contains real-time news feed with an archive from January 1, 1987 to July 31, 2017.

There are on average about 66% news articles assigned with one or more firm tags to which each article is relevant. We remove all articles with more than one firms tagged to avoid confusion, which removes another 18% articles. We collect the date, the exact timestamp, tickers of the tagged firms, the title and the content of each article.

We then match each article with the CRSP database on its tagged firm. In particular, we collect the market cap, CRSP adjusted daily returns, and the size, value, and momentum categories (according to Fama and French (1992)) to which these firms belong.

We merge each article published between 4pm of day $t-1$ and 4pm of day t with its tagged firm's current day return (close on day $t-1$ to close on day t), three-day return (close on day $t-2$ to close on day $t+1$), the next-day return (close on day t to close on day $t+1$), up to the next ten-day return, respectively. The three-day return is more robust to the simultaneous causality problem in the sense that the news might cover some events that have occurred or some that are about to occur. We use three-day returns to train and validate the model. The next-day up to the next ten-day returns are used for prediction purpose, and the current-day return for benchmark and robustness check.

Table 1 provides details on how our data filters streamline each of the subsamples, respectively. Figure 1 plots the average number of articles in each half hour throughout a day, and the average number each month over a year. The arrival of news articles is more intensive prior to US market opening and close. There is also a clear pattern on the average number of articles across different days. There is clear a leap year effect and a holiday effect on the number of news articles. The bumps around February, May, August, and November could be due to quarterly earnings, as shown in Tetlock et al. (2008) for the same dataset but for a different sampling period. Figure 3 provides a time series plot of the total number of news articles in our sample, which clearly displays an increasing trend.

Finally, we take a few steps that are common in the field of natural language processing to clean and structure the news articles.[3] The first step is normalization, including 1) changing all words in

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

Table 1: Summary Statistics

Filter	Remaining Sample Size	Observations Removed
Total Number of Dow Jones Newswire Articles	31,629,998	
Remove articles with no stocks tagged	21,128,796	10,501,202
Remove articles with more than one stocks tagged	15,209,139	5,919,657
Remove articles whose tagged stocks have no three consecutive daily returns from CRSP	10,789,892	4,419,247

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, 1987 to July 31, 2017. The last filter is only used for training and validation steps, in which we use all articles that have three-day returns matched.

the article to lower case letters; 2) expanding contractions such as "haven't" to "have not"; 3) deleting numbers, punctuations, special symbols, and only keep English words;[4] 4) combining all paragraphs into one and eliminating spaces in between. The second step is lemmatization, which groups together the different inflected forms of a word to analyze them as a single item, e.g., "growing" to "grow", etc.[5] The third step is Tokenization, which splits each article into a list of words. The fourth step is to clean stop words such as "and", "the", "is", "are", etc.[6] Finally, we translate each article into a vector of word counts, namely, its bag of words representation. Combining all articles together, we also obtain the entire dictionary of all words that appear in the database.

We also obtain a list of 2,337 negative words (Fin-Neg) and a list of 353 positive words (Fin-Pos) from the Loughran-McDonald (LM) Sentiment Word Lists, for comparison purpose.[7] Though LM lists are more business oriented, their results show that once tf-idf weights are taken into account, both LM and the Harvard IV-4 TagNeg (H4N) list are significantly associated with returns, and essentially identical in their impact.

3.2 Where do Return Predictability and Noises Come from?

Prior to presenting any empirical findings, it is perhaps useful to ponder over where the predictability comes from. Market efficiency enforces the expected return to be dominated by volatility due to unforecastable news. Our hypothesis is that the information in the news cannot be instantaneously fully reflected in changes of market prices, perhaps for reasons such as limits-to-arbitrage, rational inattention, etc. This will open up profitable investment opportunities for high frequency trading strategies.

On the other hand, daily prediction is a rather challenging task. At this frequency, fundamental variables do not change, despite investors' expectations on these variables might change rapidly.

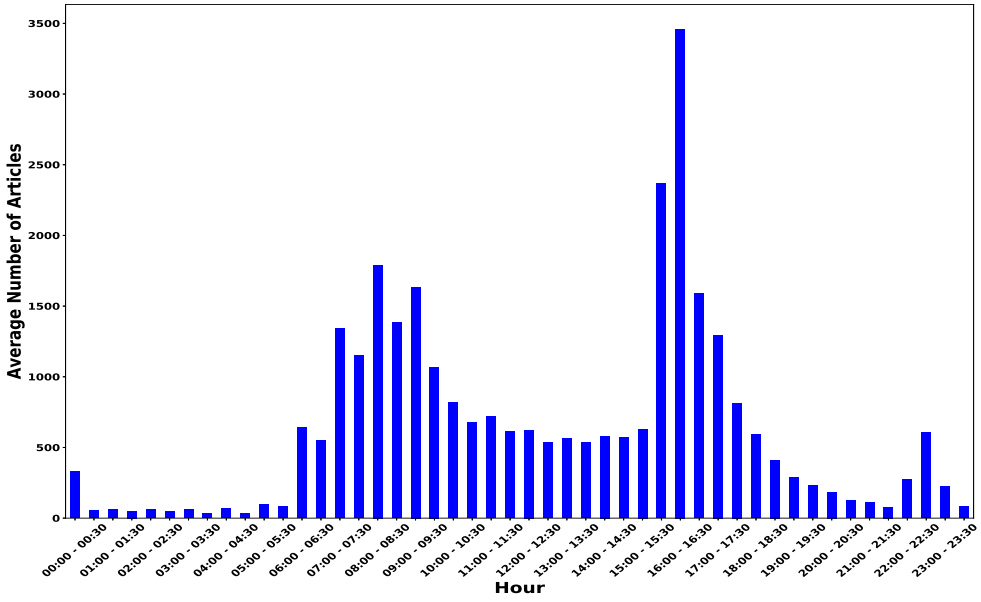
[4] The list of English words is available from item 61 on http://www.nltk.org/nltk_data/.

[5] The procedure uses the WordNet as a reference database: https://wordnet.princeton.edu/.

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

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

Figure 1: Average Number of Articles per Half an Hour



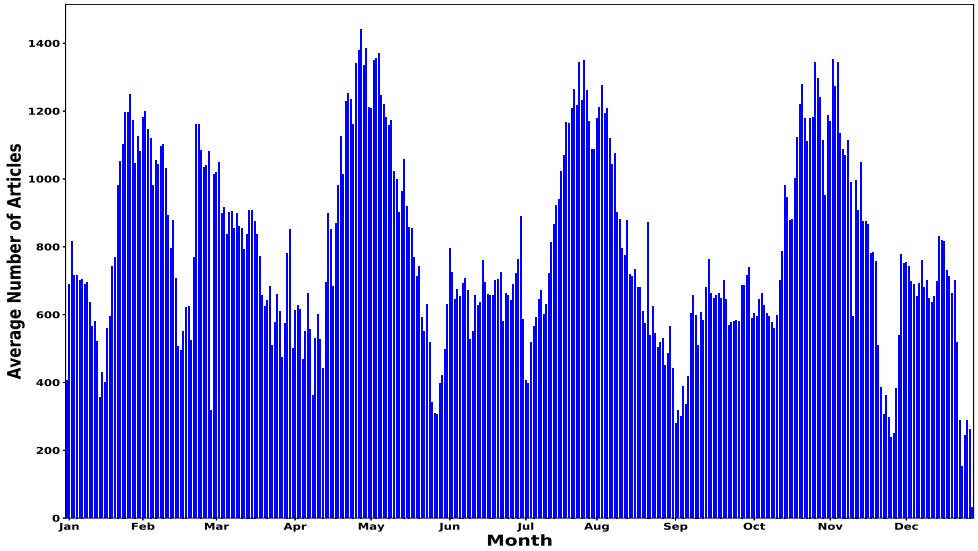
Note: This figure plots the average numbers of articles per half an hour (24 hour EST time) from January 1, 1987 to July 31, 2017.

Therefore, we have to search for information from the text data to infer investors' changes of expectations about firms' future fundamentals. There is, however, tremendous noise in the bag of words we construct from the news articles, despite that being the simplest and most common approach to text mining.

For instance, positive words are sometimes used in a negative statement, along with negation words such as "no" and "not." In addition, words like increase or decrease are ambiguous, because what these words imply depends on whether they precede words such as revenues or costs. More broadly, the bag of words ignores the context in which words appear, which often matters for their meaning. The context information is often difficult to utilize, without more complex models of words embedding. Moreover, we cannot determine the delay in the coverage of Dow Jones Newswires relative to the occurrence of the events. A specific event a news article covers might have occurred a few days or even weeks ahead in the past, in which case market efficiency has fully incorporated the information therein into returns.

That said, it is, ex-ante, unclear whether the predictability, if any, translates into profits in this low signal-to-noise ratio environment. This is precisely what motivates our empirical analysis below.

Figure 2: Average Number of Articles per Day



Note: This figure plots the average numbers of articles per day from January 1, 1987 to July 31, 2017.

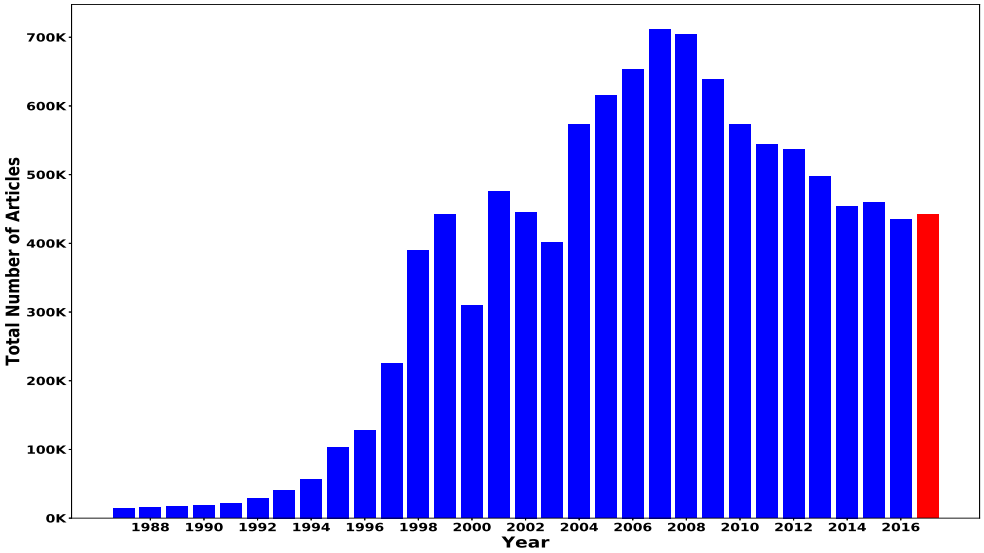
3.3 Words as Signals for Return Prediction

In this section, we go beyond regressions and directly rely on trading strategies to verify if the text data have any return predictability. Because of the large dataset, statistical significance exists even for economically insignificant effect. In fact, existing literature have discovered statistically significant associations between returns and negative words in either the HN4 or Fin-Neg dictionaries, but there exists no consensus about whether any trading strategy based on text signals leads to significant profits, once accounting for the transaction cost, see, e.g., Loughran and McDonald (2011).

To evaluate text-based trading strategies, we use models trained with a rolling sample of 10 years, validated with a 5-year rolling sample, for one year ahead out-of-sample testing. The first training sample is from 1987 to 1996, the validation sample from 1997 to 2001, and testing sample the entire year of 2002. We continue to retrain our model every year, rolling all three samples forward to include the more recent year.

In each training sample, we estimate a variety of models in each training period using a grid of tuning parameters. There are four of them, $(\alpha_+, \alpha_-, \kappa, \lambda)$. We select five choices for α_+ and α_- (25, 50, 100), 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). We then use the estimated models to score each news article in the validation sample, and select the optimal set of tuning parameters that minimizes a loss function — the ℓ^1-norm of the differences between estimated scores and the standardized ranks of all articles.

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.

In the out of sample period, we design a simple trading strategy. We sort all scores for articles each day, and long only those stocks tagged in articles that score among the top 100. Also, we build another portfolio that shorts the bottom 100 stocks. We exclude stocks with a neutral score (i.e., 0.5), if any, so the total number of stocks included may be smaller than 200, in particular for earlier years. We rebalance this portfolio on a daily basis, and use the previous day's market cap to weight all stocks. Following Tetlock et al. (2008), we exclude all articles published between 3:30pm and 4:00pm EST, which ensures traders to have at least 30 minutes, and usually much longer, to digest the information and execute the trades.

We report in Table 2 the risk-adjusted performance of these portfolios. We find that the daily returns have significant alphas against standard benchmarks in the literature. In addition, the low R^2s indicates that its return time series is not highly correlated with existing factors.

Table 2: Risk-adjusted Performance of the Constructed Portfolios

FF3		FF5		FF-Mom	
alpha	R^2	alpha	R^2	alpha	R^2
14.05%	13.01%	11.63%	15.64%	11.63%	8.22%

Note: The table report the alphas and R^2 with respect the Fama-French three-factor model ("FF3"), the Fama-French five-factor model ("FF5'), and the Fama-French five-factor model augmented to include the momentum factor ("FF5+MOM").

3.4 To What Words Returns are Most Sensitive?

We now investigate the list of words that tend to be associated with the sign of returns. It turns out the wordlist we tease out each time is remarkably stable without much variation over time. This is another advantage of using words as signals — their meanings and usage are fairly stable, compared to the technical signals used at this frequency. We provide the full list of words using the last training and validation period as an example.

Among all positive words we tease out, "undervaluing", "soar", "surge", "jump", "climb", "surpass", and "beat" consistently rank among the most optimistic words, whereas "deteriorate", "slash", "slowgoing", "warn", "downward", "blame", "plunge", "tumble", "softness", "disappoint", "plummet", "downgrade", and "shortfall" are among the most pessimistic ones. Interestingly, most of these words are verbs rather than adjectives. Also, these words seem not very often used in combination with negation in the context of news, for which succinctness and clarity are highly desirable.

Table 3: The List of Words via Screening

Negative	f_i	κ_i	Positive	f_i	κ_i
shortfall	0.272	7,045	coronary	0.549	2,190
downgrade	0.335	16,263	lift	0.548	6,725
lowerer	0.350	20,439	homecraft	0.559	3,036
disappointment	0.357	2,098	divide	0.553	7,844
disappoint	0.359	10,543	definitive	0.555	14,223
softness	0.368	4,342	thrift	0.590	2,215
tumble	0.375	2,753	upgrade	0.547	25,165
restatement	0.384	2,924	diligence	0.556	3,417
plunge	0.382	3,861	invasive	0.552	2,993
blame	0.393	5,319	mariola	0.566	2,031
downward	0.392	4,709	protatic	0.581	2,245
missel	0.396	3,063	recalcitrance	0.558	14,414
waiver	0.411	2,531	uptilt	0.559	40,028
hurt	0.394	26,800	rose	0.560	74,069
warn	0.393	14,343	authorize	0.551	21,840
slowgoing	0.398	7,034	bankbook	0.612	2,238
overambitious	0.411	2,560	beat	0.570	11,877
slash	0.405	3,312	accretive	0.569	3,978
halt	0.411	6,123	surpass	0.577	3,771
deteriorate	0.407	3,855	climb	0.581	7,285
bulletin	0.402	2,726	jump	0.586	10,016
weakness	0.408	14,564	repurchase	0.591	30,303
covenant	0.412	6,484	surge	0.584	5,590
suspend	0.422	5,288	soar	0.594	3,450
auditor	0.414	4,090	undervaluing	0.623	2,494

Note: In this table, we provide a full list of sentiment-sensitive words. The training period is from 2002 to 2011, followed by a 5-year validation period from 2012 to 2016.

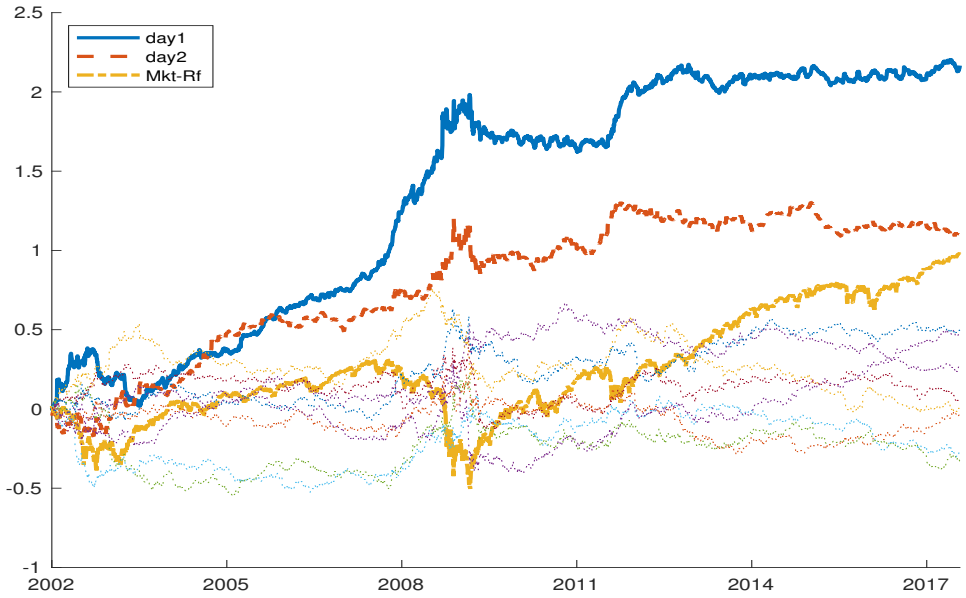
3.5 Delayed Response to News

In this section, we investigate how quickly information in the news is digested by the market. We use the same trading strategy as previously introduced, but further carry on with it up to ten working

days (two weeks). We compare the returns earned separately, on day zero (from yesterday's close to present day's close), for day one, day two, ..., and day ten. Note that the day-zero result is infeasible for trading, because it requires an existing position on stocks with news that occur afterwards. We include it only for benchmark purpose.

Figure 4 plots the cumulative returns of all portfolios, respectively. Not surprisingly, the cumulative return earned on day one is most significant, and the news impact on returns quickly disappears after few days. It is expected, perhaps because of the arrival of fresher news. In addition, a close scrutiny of the time series pattern of the cumulative returns suggests that these strategies have seen diminishing returns over time, indicating that the speed at which news are incorporated into the market has been accelerating during this period.

Figure 4: Cumulative Returns of the Constructed Portfolios



Note: Cumulative log returns of portfolios sorted on out-of-sample return forecasts. The thicker solid and dashed lines represent day-one to day-two portfolios that produce more substantial returns over this period, the thicker dot-dashed line plots the excess market return, whereas thinner lines represent portfolios of day-four and beyond, which do not show any significance.

3.6 Comparison with the LM Dictionary

We compare our framework with another approach in the literature, which uses the Fin-Neg and Fin-Pos lists, by Loughran and McDonald (2011). They show that the H4N list substantially misclassifies words when gauging tone in financial applications, and that the their Fin-Neg list is more relevant because words therein typically have negative implications in a financial sense.

We compare the cumulative returns of our long and short strategies with those based on the

Fin-Neg and Fin-Pos wordlists. We sort stocks based on their tf-idf weighted counts of positive words minus negative words, and long 100 stocks that rank on top and short 100 stocks that rank at the bottom. The cumulative returns of our long-only strategy is substantially higher. Also, the short strategy we use beats that of LM. We also find positive words are as powerful as, if not more powerful than, negative words. In fact, returns of our strategy is largely driven by the long-side of the portfolio, which is purely comprised of positive words we tease out. The gain in performance comes from the differences in our scoring strategies. First, our choice of wordlist and the weights of words are data-driven, whereas their wordlist is subjective, and their weights of words are adhoc.

3.7 Comparison with RavenPack Data

We also compare to the sentiment analysis by RavenPack. RavenPack is a leading big data analytics provider for financial services. We obtain access to their Dow Jones Edition datasets that contain their analytics based on relevant information from Dow Jones Newswires, Wall Street Journal, Barron's and MarketWatch. We collect the time-stamp, the ticker, the relevance, the composite sentiment score (CSS) of each news event in their database from Jan 1, 2002. The relevance score is between 0 and 100, indicating how strongly related the company is to the underlying news story. According to their user guide,[8] values above 75 are considered significantly relevant, so we only keep these news articles. CSS is between 1 and 100 that represents the news sentiment (with 50 indicating neutral), which is determined by emotionally charged words and phrases and by matching stories typically rated by experts as having short-term positive or negative share price impact. The score aggregates 5 sentiment analytics that combine RavenPack's Traditional, Expert Consensus, and Market Response methodologies.

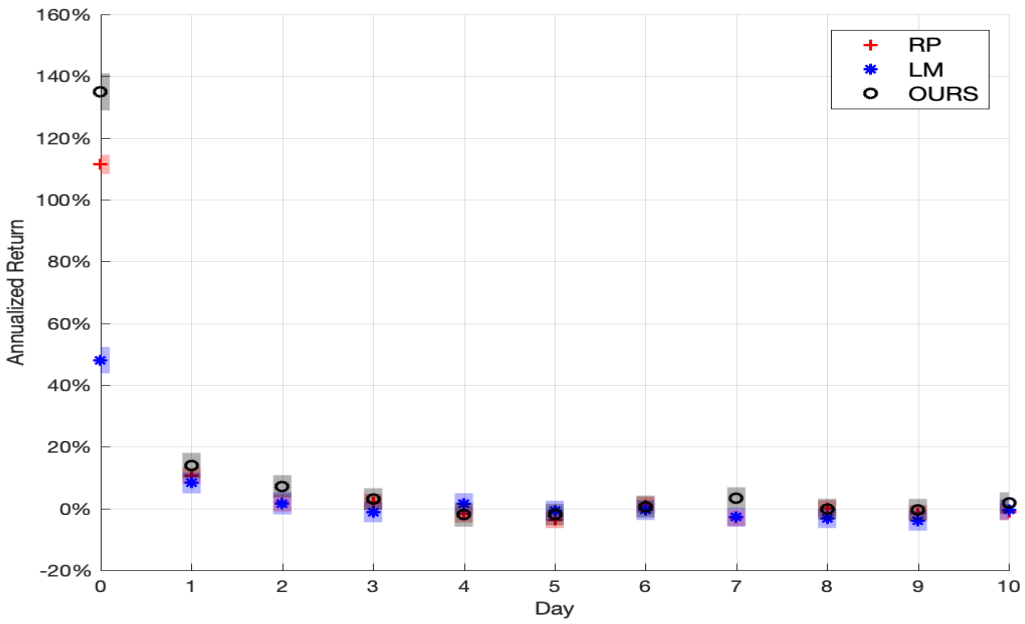
We adopt a similar strategy to build daily rebalanced portfolios by sorting on RavenPack's CSS. Although RavenPack's strategy is a black-box, the out-of-sample performance is quite similar to the strategy based on LM's dictionary. The day-zero result is much better than LM, but still dominated by our method. The day-one annualized return is on par with that based on LM. Since day two, there are no significant returns either. Because we are not informed of the exact scoring strategy used by RavenPack, the sentiment scores might have taken into account the price responses of the news, in which case the day-zero result is most suspicious to the concern of endogeneity (simultaneous causality), whereas we ensure that both LM and our method to only use past information when assigning sentiment scores to each article.

3.8 Robustness Check

We also calculate the market and sector adjusted returns, that is, residuals from CAPM and Fama French three-factor models, to check if our results are influenced by factor exposures to systemic shocks. These residuals are constructed using a rolling window regression with a sample of 252 trailing days. We use the residual returns to train and validate the model. The results, including the selected words, the scores of articles, and the out-of-sample portfolio performance, are very similar.

[8] https://ravenpack.com/rpna/newsanalytics/4.0/doc/RPNewsAnalytics_UserGuide_4.0.pdf

Figure 5: Comparison among Our Approach, LM, and RavenPack



Note: This figure compares the annualized returns of day-zero, day-one, day-two, ..., day-ten portfolios, separately. The day-zero return is not feasible for trading, as it requires an existing position on stocks prior to the occurrence of news. The day-one annualized return is conservative and feasible, barring transaction cost. The shaded area around the point estimates highlights the confidence bands.

4 Conclusion

In this paper, we illustrate how white-box machine learning tools can be applied to explore useful information from news articles for daily return prediction — a rather challenging task which only technique indicators were useful to tackle from anecdotal evidence. Our predictors are purely based on news and clearly sentiment driven. They are also interpretable, unlike technical indicators such as autocorrelations, moving averages, etc. Our empirical results show that the news driven predictability disappears quickly after few days. That said, our strategy with delay less than a day still delivers superior out of sample performance than the alternatives based on a subjective dictionary or commercial black box methods. Our trading strategy is conservative, in that in all cases the market has at least 30 minutes to digest the news information. We do not consider intraday returns, which would allow for more aggressive estimates. In addition, we do not take into account transaction cost or market impact. We leave these extensions to future work.

References

Anandkumar, Anima, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-Kai Liu, 2012, A spectral algorithm for latent dirichlet allocation, in *Advances in Neural Information Processing*

Systems, 917–925.

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, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu, 2013, A practical algorithm for topic modeling with provable guarantees, in *International Conference on Machine Learning*, 280–288.

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

Fama, Eugene F, and Kenneth R French, 1992, The cross-section of expected stock returns, *the Journal of Finance* 47, 427–465.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix

A Statistical Theory

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

$$\widehat{S} = \big\{j : \max\{f_j, 1 - f_j\} \geq 1/2 + \alpha_{\pm}\big\}, \tag{A.1}$$

where

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

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

A.1 Regularity Conditions

Let $s_{\max}$, $s_{\min}$, and $\bar{s}$ be the maximum, minimum, and average of $\{s_i\}_{i=1}^n$, respectively. In our model, for sentiment-neutral words, $d_{i,[N]}$ follows a multinomial distribution. Define $\Omega_i = \mathbb{E}d_{i,[N]}$.[9] For each $j \in N$, let $\Omega_{\min,j}$, $\Omega_{\max,j}$, and $\bar{\Omega}_{\cdot,j}$ be the maximum, minimum, and average of $\{\Omega_{i,j}\}_{i=1}^n$, respectively. We assume

$$\frac{s_{\max}}{\bar{s}} \leq C, \qquad \max_{j\in N} \frac{\Omega_{\max,j}}{\bar{\Omega}_{\cdot,j}} \leq C, \qquad \min_{j\in S} \frac{n\bar{s}(O_{+,j} + O_{-,j})}{\log(m)} \to \infty, \qquad \min_{j\in N} \frac{n\bar{\Omega}_{\cdot,j}}{\log(m)} \to \infty. \tag{A.2}$$

The last two inequalities in (A.2) require the expected count of any word in all of n training articles to be much larger than $\log(m)$. Since n is large in real data, this condition is mild. For a constant $c_0 \in (0, 1)$, we assume

$$\min_{j\in S} \frac{\sum_{i=1}^n \big[p_i O_{+,j} + (1 - p_i)O_{-,j}\big]}{\sum_{i=1}^n (O_{+,j} + O_{-,j})} \geq c_0. \tag{A.3}$$

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

$$\frac{1}{n}\sum_{i=1}^n p_i = \frac{1}{2}, \qquad \frac{\sum_{i=1}^n s_i\, \mathbb{E}[\mathrm{sgn}(y_i)]}{\sum_{i=1}^n s_i} = 0, \tag{A.4}$$

This condition essentially requires that we have approximately equal number of articles with positive and negative tone. Note that we can always keep the same number of articles associated with positive

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

and negative returns in the training stage, so this condition is mild. We also assume

$$\frac{\sum_{i=1}^n \Omega_{i,j}\,\mathbb{E}[\mathrm{sgn}(y_i)]}{\sum_{i=1}^n \Omega_{i,j}} = 0, \qquad \text{for all } j \in N. \tag{A.5}$$

This condition ensures that the count of any sentiment-neutral word has no correlation with the sign of the stock returns (so they are indeed "sentiment-neutral"). All equalities in (A.4)-(A.5) do not need to hold exactly. We impose exact inequalities so that the conclusions are more elegant.

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

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

$$\theta \equiv \frac{\sum_{i=1}^n s_i\big(p_i - \frac{1}{2}\big)\big[g(p_i) - \frac{1}{2}\big]}{\sum_{i=1}^n s_i}. \tag{A.6}$$

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

Theorem A.1. *Consider the model (1)-(4), where (A.2)-(A.5) hold. As $n, m \to \infty$, with probability $1 - o(1)$,*

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

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

Theorem A.2 (Sure Screening)**.** *Consider the model (1)-(4), where (A.2)-(A.5) hold. We assume*

$$n\theta^2 \min_{j\in S} \frac{(O_{+,j} - O_{-,j})^2}{(O_{+,j} + O_{-,j})^2} \geq \frac{\log^2(m)}{\min\big\{1,\ \bar{s}\min_{j\in S}(O_{+,j} + O_{-,j}),\ \min_{j\in N}\bar{\Omega}_{\cdot,j}\big\}}. \tag{A.7}$$

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

The desired number of training articles for sure screening is determined by three factors. First, θ. The sensitivity of stock returns to article polarity, defined in (A.6). Second, $\min_{j\in S}\frac{|O_{+,j}-O_{-,j}|}{O_{+,j}+O_{-,j}}$. It represents the word's frequency-adjusted sentiment. Third, $\min\{1, \bar{s}\min_{j\in S}(O_{+,j}+O_{-,j}), \min_{j\in N}\bar{\Omega}_{\cdot,j}\}$. Note that the last two terms in the minimum are related to the per-article count of individual words. For "long articles" where the per-article count of each word is bounded below by a constant, this factor equals 1. For "short articles", the per-article count of a word may tend to zero, so we need to have more training articles.

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

Theorem A.3 (Estimation Error of Sentiment Vectors)**.** *Consider the model* (1)-(4)*, where* (A.2)-(A.5) *and* (A.7) *hold. With probability* $1 - o(1)$*,*

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

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

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

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

However, the supervised approach comes with a price: Our method is estimating $(F, \rho T)$, instead of (F, T). Note that our assumption (2) ensures $\rho > 0$. It means, regardless of the errors of estimating p_i by $\widetilde{p}_i$, our method always *preserves the order of the tone of words*. This property is very important, as it guarantees that in the scoring step we always correctly identify whether a new article has positive or negative sentiment, regardless of the errors in $\widetilde{p}_i$.

When $\widetilde{p}_i = p_i$, the factor $\rho = 1$.[10] So, our method precisely estimates T. When $p_i \neq p_i$, this factor is smaller than 1, so our method "discounts" the vector of tone. Once the exact distribution of y_i given p_i is specified, this factor can be computed explicitly. In the special case where p_i is uniformly distributed on $[0, 1]$ and $y_i \sim N(p_i, \sigma^2)$, we have calculated the value ρ as $n \to \infty$:[11]

σ	0.001	0.01	0.05	0.1
ρ	1.0000	0.9994	0.9861	0.9491

When $\sigma = 0.1$, there is a huge error of using $\widetilde{p}_i$ to estimate p_i, but ρ is still close to 1.

[10] By the way of construction, $\widetilde{p}_i$ is uniformly distribution on $[0, 1]$. So, when $p_i = \widetilde{p}_i$, we have $\rho = 12\int_0^1 (x - \frac{1}{2})^2 dx = 1$.

[11] The numbers are numerically calculated using 20 repetitions at $n = 1,000,000$.

A.3 Accuracy of the Estimator in Algorithm 3

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

$$p^* = \frac{1}{2} + \rho^{-1}\Big(p - \frac{1}{2}\Big). \tag{A.8}$$

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

Theorem A.4 (Scoring Error on New Article)**.** *Consider the model (1)-(4), where (A.2)-(A.5) hold. Define $O^{(\rho)} = [O_+^{(\rho)}, O_-^{(\rho)}]$, with $O_\pm^{(\rho)} = F \pm \rho T$. Suppose (A.7) is satisfies with O replaced by $O^{(\rho)}$. Let $d \in \mathbb{R}_+^m$ be the word count vector of a new article with polarity p. For a constant $c_1 \in (0, \frac{1}{2})$, we assume that $pO_{+,j} + (1-p)O_{-,j} \geq c_1(O_{+,j} + O_{-,j})$, for all $j \in S$, and that $c_1 \leq p^* \leq 1 - c_1$, where p^* is the rescaled polarity. Write*

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

We assume the length of the new article satisfies $s\Theta \to \infty$. Let $\widehat{p}$ be the estimator in (12) with a tuning parameter $\lambda > 0$. For any $\epsilon > 0$, with probability $1 - \epsilon$,

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

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

We make a few remarks. Our scoring step estimates the rescaled polarity p^*, instead of the true p. Since $(p^* - \frac{1}{2})$ and $(p - \frac{1}{2})$ have the same sign, our method correctly identifies whether the article has positive or negative sentiment, which is the most critical information for constructing the trading strategies. Furthermore, if $\widetilde{p}_i \approx p_i$ is small in the training stage, then $p^* \approx p$, and our method also correctly estimates p.

The choice of λ yields a bias-variance trade-off. In the error bound for $|\widehat{p} - p^*|$, the first term $\min\{1, \frac{\rho^2\Theta}{\lambda}\}err_n$ is the "variance" term, decreasing with λ; the second term $\min\{1, \frac{\lambda}{\rho^2\Theta}\}|p^* - \frac{1}{2}|$ is the "bias" term, increasing with λ. In reality, it is a common belief that the majority of articles have a neutral tone, so the bias is negligible. At the same time, text data are very noisy, so adding the penalty can significantly reduce the variance. Our estimator shares the same spirit as the James-Stein estimator (James and Stein, 1961) by shrinking the MLE of p towards $\frac{1}{2}$. Interestingly, given that the true polarity p is closer to $\frac{1}{2}$ than p^*, the shrinkage effect here helps reduce the scaling effect in (A.8), which means in some scenarios our estimator does a better job estimating the original p.

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

B Mathematical Proofs

B.1 Proofs of Theorem A.1 and Theorem A.2

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

$$f_j = \frac{1}{2} + \frac{1}{2}\frac{L_j^+ - L_j^-}{L_j^+ + L_j^-} = \frac{1}{2} + \frac{\sum_{i=1}^n t_i \cdot d_{i,j}}{\sum_{i=1}^n d_{i,j}}. \tag{B.9}$$

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

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

$$d_{i,j} \sim \mathrm{Binomial}\big(s_i,\ F_j + \eta_i T_j\big). \tag{B.10}$$

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

$$f_j \overset{(d)}{=} \frac{1}{2} + \frac{\sum_{i=1}^n \sum_{\ell=1}^{s_i} t_i \cdot b_{i,j,\ell}}{\sum_{i=1}^n \sum_{\ell=1}^{s_i} b_{i,j,\ell}}, \qquad \text{where} \quad b_{i,j,\ell} \overset{iid}{\sim} \mathrm{Bernoulli}(F_j + \eta_i T_j). \tag{B.11}$$

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

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

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

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

We plug the above inequalities into (B.11). It gives

$$f_j = \frac{1}{2} + \frac{\sum_{i=1}^n t_i s_i (F_j + \eta_i T_j) + O\big(\sqrt{n\bar{s} F_j \log(m)}\big)}{\sum_{i=1}^n s_i (F_j + \eta_i T_j) + O\big(\sqrt{n\bar{s} F_j \log(m)}\big)}$$
$$= \frac{1}{2} + \frac{F_j \sum_{i=1}^n t_i s_i + T_j \sum_{i=1}^n t_i \eta_i s_i + O\big(\sqrt{n\bar{s} F_j \log(m)}\big)}{F_j \sum_{i=1}^n s_i + T_j \sum_{i=1}^n \eta_i s_i + O\big(\sqrt{n\bar{s} F_j \log(m)}\big)}. \tag{B.12}$$

In the denominator, the sum of the first two terms can be rewritten as $\sum_{i=1}^n s_i[p_i O_{+,j} + (1-p_i)O_{-,j}]$. It is upper bounded by $2n\bar{s}F_j$, and by (A.3), it is also lower bounded by $2c_0 n\bar{s}F_j$. Furthermore, since $n\bar{s}F_j \gg \log(m)$, the last term is negligible compared to the first two terms. Hence, the denominator in (B.12) is between $c_0 n\bar{s}F_j$ and $4n\bar{s}F_j$. It follows that

$$|f_j - 1/2| \geq \frac{|T_j \sum_{i=1}^n t_i \eta_i s_i|}{4n\bar{s}F_j} - \frac{|F_j \sum_{i=1}^n t_i s_i|}{c_0 n\bar{s}F_j} + \frac{O\big(\sqrt{n\bar{s}F_j \log(m)}\big)}{c_0 n\bar{s}F_j} \tag{B.13}$$

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

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

In particular, we know that $|\sum_{i=1}^n \eta_i s_i t_i| \geq 2n\bar{s}\theta$. Similarly, with probability $1 - O(m^{-2})$, $|\sum_{i=1}^n s_i t_i - \sum_{i=1}^n s_i \mathbb{E}t_i| \leq C\bar{s}\sqrt{n\log(m)}$. Note that $\sum_{i=1}^n s_i \mathbb{E}t_i = 0$, due to the second equality in (A.4). So, we have $|\sum_{i=1}^n s_i t_i| \leq C\bar{s}\sqrt{n\log(m)}$. We plug these results into (B.13) and find out that

$$|f_j - 1/2| \geq \frac{|T_j| 2n\bar{s}\theta}{4n\bar{s}F_j} - \frac{F_j \cdot C\bar{s}\sqrt{n\log(m)}}{c_0 n\bar{s}F_j} + \frac{O\big(\sqrt{n\bar{s}F_j\log(m)}\big)}{c_0 n\bar{s}F_j}$$
$$\geq \frac{\theta|T_j|}{2F_j} + O\Big(\sqrt{\tfrac{\log(m)}{n}}\Big) + O\Big(\sqrt{\tfrac{\log(m)}{n\bar{s}F_j}}\Big). \tag{B.14}$$

This gives the first claim of Theorem A.1.

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

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

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

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

Plugging them into (B.9) gives

$$\begin{aligned} f_j &= \frac{1}{2} + \frac{\sum_{i=1}^n t_i\Omega_{i,j} + O\big([n\bar{\Omega}_{\cdot,j}\log(m)]^{\frac{1}{2}}\big)}{\sum_{i=1}^n \Omega_{i,j} + O\big([n\bar{\Omega}_{\cdot,j}\log(m)]^{\frac{1}{2}}\big)} \\ &= \frac{1}{2} + \frac{\sum_{i=1}^n t_i\Omega_{i,j} + O\big([n\bar{\Omega}_{\cdot,j}\log(m)]^{\frac{1}{2}}\big)}{n\bar{\Omega}_{\cdot,j} + O\big([n\bar{\Omega}_{\cdot,j}\log(m)]^{\frac{1}{2}}\big)}. \end{aligned} \tag{B.15}$$

We then deal with the randomness of $\{t_i\}_{i=1}^n$. By Hoeffding's inequality, with probability $1-O(m^{-2})$, $|\sum_{i=1}^n \Omega_{i,j}(t_i - \mathbb{E}t_i)| \le C\sqrt{\sum_{i=1}^n \Omega_{i,j}^2 \log(m)} \le C\bar{\Omega}_{\cdot,j}\sqrt{n\log(m)}$, where the last inequality is from the condition $\Omega_{\max,j} \le C\bar{\Omega}_{\cdot,j}$. Moreover, by our condition (A.4), $\sum_{i=1}^n \Omega_{i,j}\mathbb{E}t_i = 0$. The above imply

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

We plug it into (B.15) and note that the denominator of (B.15) is $\gtrsim n\bar{\Omega}_{\cdot,j}$, since $n\bar{\Omega}_{\cdot,j} \gg \log(m)$. It follows that

$$\begin{aligned} |f_j - 1/2| &\le \frac{C\bar{\Omega}_{\cdot,j}\sqrt{n\log(m)} + O\big([n\bar{\Omega}_{\cdot,j}\log(m)]^{\frac{1}{2}}\big)}{n\bar{\Omega}_{\cdot,j}} \\ &\le O\Big(\sqrt{\tfrac{\log(m)}{n}}\Big) + O\Big(\sqrt{\tfrac{\log(m)}{n\bar{\Omega}_{\cdot,j}}}\Big). \end{aligned} \tag{B.16}$$

This gives the second claim of Theorem A.1.

Next, we prove Theorem A.2. By (B.14) and (B.16), with probability $1-O(m^{-1})$, simultaneously for all $1 \le j \le m$,

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

where $e_n^2 = (\min\{1, \bar{s}\min_{j\in S} F_j, \min_{j\in N}\bar{\Omega}_{\cdot,j}\})^{-1}\frac{\log(m)}{n}$. The assumption (A.7) ensures that $\frac{\theta|T_j|}{2F_j} \gg e_n\sqrt{\log(m)}$. By setting the threshold at $e_n\sqrt{\log(\log(m))}$, all words in S will retain and all words in N will be screened out. □

B.2 Proof of Theorem A.3

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

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

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

$$\|F^* - F\|_1 = O(n^{-1}), \qquad \|T^* - \rho T\|_1 = O(n^{-1}) \tag{B.17}$$

In the second part of our proof, we show that

$$\|\widehat{O}_\pm - O^*_\pm\|_1 \leq C\sqrt{|S|\log(m)/(n\bar{s})}. \tag{B.18}$$

The claim follows by combining (B.17)-(B.18).

First, we show (B.17). By definition,

$$\begin{aligned}[F^*, T^*] = O^* \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} &= O(W\widetilde{W})(\widetilde{W}\widetilde{W}')^{-1} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \\ &= [F, T] \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} (W\widetilde{W})(\widetilde{W}\widetilde{W}')^{-1} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}}_{\equiv M}. \end{aligned} \tag{B.19}$$

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

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

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 $\widetilde{W}\widetilde{W}'$: First, $\sum_{i=1}^n \widetilde{p}_{(i)}^2 = \frac{1}{n^2}\sum_{i=1}^n i^2 = \frac{n}{3}[1+O(n^{-1})]$. Second, $\sum_{i=1}^n (1-\widetilde{p}_{(i)})\widetilde{p}_{(i)} = \frac{1}{n^2}\sum_{i=1}^n i(n-i) = \frac{1}{n}\sum_{i=1}^n i - \frac{1}{n^2}\sum_{i=1}^n i^2 = \frac{n}{6}[1+O(n^{-1})]$. Third, $\sum_{i=1}^n (1-\widetilde{p}_{(i)})^2 = \frac{1}{n^2}\sum_{i=1}^n (n-i)^2 = \frac{1}{n^2}\sum_{i=0}^{n-1} i^2 = \frac{n}{3}[1+O(n^{-1})]$. Combining them gives

$$n^{-1}(\widetilde{W}\widetilde{W}') = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{bmatrix} + O(n^{-1}) \quad \Longrightarrow \quad n(\widetilde{W}\widetilde{W}')^{-1} = \begin{bmatrix} 4 & -2 \\ -2 & 4 \end{bmatrix} + O(n^{-1}). \tag{B.20}$$

Additionally, by direct calculations,

$$n^{-1}(W\widetilde{W}') = \begin{bmatrix} \frac{1}{n}\sum_i p_i\widetilde{p}_i & \frac{1}{n}\sum_i p_i(1-\widetilde{p}_i) \\ \frac{1}{n}\sum_i (1-p_i)\widetilde{p}_i & \frac{1}{n}\sum_i (1-p_i)(1-\widetilde{p}_i) \end{bmatrix} \tag{B.21}$$

We now plug (B.20)-(B.21) into (B.19). It gives

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

The condition (A.4) yields $M_{21} = 0$. The way we construct $\{\widetilde{p}_i\}_{i=1}^n$ ensures $M_{12} = O(n^{-1})$. Combined with the definition of ρ in (11), the above imply

$$M = \begin{bmatrix} 1 & 0 \\ 0 & \rho \end{bmatrix} + O(n^{-1}). \tag{B.22}$$

Then, (B.17) follows from plugging in (B.22) into (B.19).

Second, we show (B.18). Let $\overline{O} = [\overline{O}_+, \overline{O}_-]$ be the matrix obtained from setting negative entries of $\widetilde{O}$ to zero. Algorithm 2 outputs $\widehat{O}_\pm = (1/\|\overline{O}_\pm\|_1)\overline{O}_\pm$. It follows that, for $j \in S$,

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

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

$$|\widehat{O}_{\pm,j} - O^*_{\pm,j}| \le |\overline{O}_{\pm,j} - O^*_{\pm,j}| + \frac{|\overline{O}_{\pm,j}|}{\|\overline{O}_\pm\|_1}\|\overline{O}_\pm - O^*_\pm\|_1. \tag{B.23}$$

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

$$\|\widehat{O}_\pm - O^*_\pm\|_1 \le 2\|\widetilde{O}_\pm - O^*_\pm\|_1. \tag{B.24}$$

Therefore, to show (B.18), it suffices to bound $\|\widetilde{O}_\pm - O^*_\pm\|_1$.

Let W be the matrix whose i-th column is $(p_i, 1-p_i)'$. By model (4), $s_i\widetilde{d}_i \sim \text{Multinomial}(s_i, p_iO_+ + (1-p_i)O_-)$. It leads to $\mathbb{E}\widetilde{d}_i = (OW)_i$. Write $Z = \widetilde{D} - \mathbb{E}\widetilde{D}$. Then, $\widetilde{D} = OW + Z$ and

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

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

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

It follows that

$$\begin{aligned}\|\widetilde{O}_{\pm,j} - O^*_{\pm,j}\|_1 &\le \max\Big\{\Big|\frac{1}{n}\sum_{i=1}^n \widetilde{p}_i Z_{i,j}\Big|,\ \Big|\frac{1}{n}\sum_{i=1}^n (1-\widetilde{p}_i)Z_{i,j}\Big|\Big\}\big\|n(\widetilde{W}\widetilde{W}')^{-1}\big\|_1 \\ &\le C\max\Big\{\Big|\frac{1}{n}\sum_{i=1}^n \widetilde{p}_i Z_{i,j}\Big|,\ \Big|\frac{1}{n}\sum_{i=1}^n (1-\widetilde{p}_i)Z_{i,j}\Big|\Big\},\end{aligned} \tag{B.25}$$

where in the last line we have used (B.21). We now bound $|\frac{1}{n}\sum_{i=1}^n \widetilde{p}_i Z_{i,j}|$. The bound for $|\frac{1}{n}\sum_{i=1}^n (1-\widetilde{p}_i)Z_{i,j}|$ can be obtained similarly, so the proof is omitted. Since $\{\widetilde{p}_i\}_{i=1}^n$ are constructed from $\{y_i\}_{i=1}^n$, they are independent of $\{Z_{i,j}\}_{i=1}^n$ by our assumption (1). We thus condition on $\{\widetilde{p}_i\}_{i=1}^n$. Let $\{b_{i,j,\ell}\}_{\ell=1}^{s_i}$

be a collection of *iid* Bernoulli variables with a success probability $[p_i O_{+,j} + (1-p_i)O_{-,j}]$. Then, $d_{i,j}$ has the same distribution as $\sum_{\ell=1}^{s_i} b_{i,j,\ell}$. It follows that $Z_{i,j} \stackrel{(d)}{=} \sum_{\ell=1}^{s_i} s_i^{-1}(b_{i,j,\ell} - \mathbb{E}b_{i,j,\ell})$. Hence,

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

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

$$\Big|\sum_{i=1}^{n} \widetilde{p}_i Z_{i,j}\Big| \le C\sqrt{n\bar{s}^{-1}F_j \log(m)} + C\bar{s}^{-1}\log(m) \le C\sqrt{n\bar{s}^{-1}F_j \log(m)}, \tag{B.26}$$

where the last line is due to $n\bar{s}F_j/\log(m) \to \infty$. The bound for $|\sum_{i=1}^n (1-\widetilde{p}_i)Z_{i,j}|$ is similar. Plugging them into (B.25) gives

$$\|\widetilde{O}_{\pm,j} - O^*_{\pm,j}\|_1 \le C\frac{\sqrt{F_j \log(m)}}{\sqrt{n\bar{s}}}. \tag{B.27}$$

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.18). The proof is now complete. □

B.3 Proof of Theorem A.4

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

We need some preparation. First, by our assumption, $F_j + \eta T_j = pO_{+,j} + (1-p)O_{-,j} \ge c_1(O_{+,j} + O_{-,j}) = 2c_1 F_j$. Second, by (B.27) in the proof of Theorem A.3, $|\widehat{F}_j - F_j| \le C\sqrt{F_j\log(m)/(n\bar{s})}$ and $|\widehat{T}_j - \rho T_j| \le C\sqrt{F_j\log(m)/(n\bar{s})}$. Since $n\bar{s}F_j \gg \log(m)$, we immediately obtain $|\widehat{F}_j - F_j| = o(F_j)$. Third, the condition (A.7) guarantees $n\theta^2\frac{\rho^2 T_j^2}{F_j^2} \ge \frac{\log^2(m)}{\bar{s}F_j}$. In other words, $\rho|T_j| \gg \sqrt{F_j\log(m)/(n\bar{s})}$. So, $|\widehat{T}_j - \rho T_j| \ll \rho|T_j|$. We summarize these results as follows: for any $j \in S$,

$$\frac{|F_j + \eta T_j|}{F_j} \ge 2c_1, \qquad \frac{\max\{|\widehat{F}_j - F_j|, |\widehat{T}_j - \rho T_j|\}}{F_j} \le C\sqrt{\frac{\log(m)}{n\bar{s}F_j}}, \qquad \frac{|\widehat{T}_j - \rho T_j|}{\rho|T_j|} = o(1). \tag{B.28}$$

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

$$|\widehat{p} - p^*| = \frac{1}{2}|\widehat{\eta} - \rho^{-1}\eta|. \tag{B.29}$$

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

Suppose $|\widehat{\eta}-\rho^{-1}\eta| \le c_1$. Let $\widehat{F} = \frac{1}{2}(\widehat{O}_+ + \widehat{O}_-)$ and $\widehat{T} = \frac{1}{2}(\widehat{O}_+ - \widehat{O}_-)$. Since $p(1-p) = (1-\eta^2)/4$ and $p\widehat{O}_{+,j} + (1-p)\widehat{O}_{-,j} = \widehat{F} + \eta\widehat{T}_j$, the penalized MLE (12) has an equivalent form:

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

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

$$s^{-1}\sum_{j\in S} d_j \log\Big(1 + \frac{(\widehat{\eta}-\rho^{-1}\eta)\widehat{T}_j}{\widehat{F}_j + \rho^{-1}\eta\widehat{T}_j}\Big) + \lambda\log\Big(1 + \frac{\widehat{\eta}-\rho^{-1}\eta}{1+\rho^{-1}\eta}\Big) + \lambda\log\Big(1 - \frac{\widehat{\eta}-\rho^{-1}\eta}{1-\rho^{-1}\eta}\Big) \ge 0. \tag{B.30}$$

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

$$\begin{aligned}
&\log\Big(1 + \frac{\widehat{\eta}-\rho^{-1}\eta}{1+\rho^{-1}\eta}\Big) + \log\Big(1 - \frac{\widehat{\eta}-\rho^{-1}\eta}{1-\rho^{-1}\eta}\Big) \\
\le\ & \frac{\widehat{\eta}-\rho^{-1}\eta}{1+\rho^{-1}\eta} - \frac{(\widehat{\eta}-\rho^{-1}\eta)^2}{4(1+\rho^{-1}\eta)^2} - \frac{\widehat{\eta}-\rho^{-1}\eta}{1-\rho^{-1}\eta} - \frac{(\widehat{\eta}-\rho^{-1}\eta)^2}{4(1-\rho^{-1}\eta)^2} \\
=\ & -(\widehat{\eta}-\rho^{-1}\eta)\frac{2\rho^{-1}\eta}{1-\rho^{-2}\eta^2} - (\widehat{\eta}-\rho^{-1}\eta)^2\frac{1+\rho^{-2}\eta^2}{2(1-\rho^{-2}\eta^2)^2}.
\end{aligned} \tag{B.31}$$

Also, by (B.28), $\widehat{F}_j + \rho^{-1}\eta\widehat{T}_j \sim F_j + \eta T_j \ge 2c_1 F_j$ and $|\widehat{T}_j| \sim \rho|T_j|$. Hence, $\big|\frac{(\widehat{\eta}-\rho^{-1}\eta)\widehat{T}_j}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j}\big| \le |\widehat{\eta}-\rho^{-1}\eta|\cdot\frac{\rho}{2c_1}$, which is bounded by $\frac{1}{2}$ on the event $|\widehat{\eta}-\rho^{-1}\eta| \le c_1$. Note that $\log(1+x) \le x - \frac{x^2}{4}$ for $x \in [-\frac{1}{2}, \frac{1}{2}]$. We thus have

$$\begin{aligned}
&s^{-1}\sum_{j\in S} d_j \log\Big(1 + \frac{(\widehat{\eta}-\rho^{-1}\eta)\widehat{T}_j}{\widehat{F}_j + \rho^{-1}\eta\widehat{T}_j}\Big) \\
\le\ & (\widehat{\eta}-\rho^{-1}\eta)\sum_{j\in S}\frac{s^{-1}d_j\widehat{T}_j}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j} - (\widehat{\eta}-\rho^{-1}\eta)^2\sum_{j\in S}\frac{s^{-1}d_j\widehat{T}_j^2}{4(\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j)^2}.
\end{aligned} \tag{B.32}$$

We plug (B.31)-(B.32) into (B.30). It gives

$$(\widehat{\eta}-\rho^{-1}\eta)X_1 - (\widehat{\eta}-\rho^{-1}\eta)^2X_2 \ge 0, \qquad \Longrightarrow \qquad |\widehat{\eta}-\rho^{-1}\eta| \le \frac{|X_1|}{X_2}, \tag{B.33}$$

where

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

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

Consider X_1. Since $(\widehat{F}, \widehat{T})$ are obtained from the training data, they are independent of d. We

thus condition on $(\widehat{F}, \widehat{T})$. Using (B.28), we can get

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

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

$$\begin{aligned}
\mathrm{Var}\left(\sum_{j\in S}\frac{\widehat{T}_j s^{-1}d_j}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j}\right)&\leq\sum_{j\in S}\frac{\widehat{T}_j^2 s^{-2}\cdot s(F_j+\eta T_j)}{(\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j)^2} \\
&\leq Cs^{-1}\sum_{j\in S}\frac{\rho^2T_j^2}{F_j}+Cs^{-1}\sum_{j\in S}\frac{(\widehat{T}_j-\rho T_j)^2}{F_j} \\
&\leq Cs^{-1}\rho^2\Theta+Cs^{-1}\frac{|S|\log(m)}{n\bar{s}},
\end{aligned}$$

where we have used (B.28). Let $\{b_\ell\}_{\ell=1}^s$ be *iid* variables, where $b_\ell\sim\mathrm{Multinomial}(1,F+\eta T)$. Then, d has the same distribution as $\sum_{\ell=1}^s b_\ell$. It follows that

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

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

$$\begin{aligned}
\left|\sum_{j\in S}\frac{\widehat{T}_j s^{-1}d_j}{\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j}\right|&\leq C\sqrt{\frac{|S|\log(m)}{n\bar{s}}}+C\rho\sqrt{\frac{\Theta\log(\epsilon^{-1})}{s}}+\frac{\rho\log(\epsilon^{-1})}{2c_1 s} \\
&\leq C\sqrt{\frac{|S|\log(m)}{n\bar{s}}}+C\rho\sqrt{\frac{\Theta\log(\epsilon^{-1})}{s}},
\end{aligned}\tag{B.34}$$

where the last line is because $s\Theta\to\infty$. We plug (B.34) into the expression of X_1. Additionally, we

notice that $1-\rho^{-2}\eta^{-2} = (1+\rho^{-1}\eta)(1-\rho^{-1}\eta) = 4p^*(1-p^*) \geq 4c_1^2$. Hence, with probability $1-\epsilon$,

$$|X_1| \leq \frac{\lambda}{2c_1^2}|\rho^{-1}\eta| + C\sqrt{\frac{|S|\log(m)}{n\bar{s}}} + C\rho\sqrt{\frac{\Theta\log(\epsilon^{-1})}{s}}. \tag{B.35}$$

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

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

At the same time,

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

Similarly as proving (B.34), we then introduce variables $\{b_\ell\}_{\ell=1}^s$ and apply the Bernstein's inequality. Note that the above variance is much smaller than the square of the mean, due to $s\Theta\to\infty$. It follows that, with probability $1-\epsilon$,

$$\sum_{j\in S}\frac{\widehat{T}_j^2 s^{-1}d_j}{4(\widehat{F}_j+\rho^{-1}\eta\widehat{T}_j)^2} \geq C^{-1}\rho^2\Theta. \tag{B.36}$$

We plug (B.36) into the expression of X_2 and note that $1-\rho^{-2}\eta^2 = 4p^*(1-p^*) \leq 1$. It yields that

$$X_2 \geq \frac{\lambda}{2} + C^{-1}\rho^2\Theta. \tag{B.37}$$

We now plug (B.35) and (B.37) into (B.33). It follows that

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

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

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

Combining it with (B.29) and noting that $\rho^{-1}\eta = 2(p^*-\frac{1}{2})$, we have the desired claim.

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

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

As $\eta\to+1$, $\ell_\lambda'(\eta)\to-\infty$; as $\eta\to-1$, $\ell_\lambda'(\eta)\to+\infty$. Hence, the maximum is attained in the interior

of $(-1,1)$. Since the true $p^* \in [c_1, 1-c_1]$, it follows that $|\rho^{-1}\eta| \le |1-2c_1|$. We now evaluate $\ell'_\lambda(\cdot)$ at $1-1.9c_1$. Following the same argument as proving (B.34), we can show that

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

So, it is strictly negative. As a result, the maximum cannot be attained at $[1-1.9c_1, 1)$. Similarly, we can prove that the maximum cannot be attained at $(-1, -1+1.9c_1]$. Now, we have restricted our attention to a compact interval that is bounded away from ± 1 by at least $1.9c_1$. For any η_0 in this interval, $F_j + \eta_0 T_j \ge cF_j$ for a constant $c > 0$. This allows us to mimic the proof of (B.36)-(B.37) to get

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

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

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

If $|\widehat{\eta} - \rho^{-1}\eta| > c_1$, then the above implies $|\ell'_\lambda(\rho^{-1}\eta)| \ge c_1|\ell''_\lambda(\eta_0)| \ge C^{-1}(\lambda+\rho^2\Theta)$. On the other hand, we notice that $X_1 = \ell'_\lambda(\rho^{-1}\eta)$, where we have proved in (B.35) that $|X_1| = o(\lambda+\rho^2\Theta)$. This yields a contradiction. The proof is now complete. □