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

February 13, 2019

Abstract

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

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

The first step 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 and low signal-to-noise ratios.

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 a 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. 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.

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 opposed to traditional firm fundamental data used predominantly in the academic finance literature. Also, our paper focuses on a daily prediction horizon. Known predictors for daily returns in the existing literature are typically technical signals, constructed from the observed price path, the associated volumes, their moving averages, etc. These technical indicators are difficult to interpret. In contrast, interpretability is self-evident for text data.

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 a list of positive and negative words has 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 quite substantial, thanks to the use of more powerful machine learning techniques.

There are few exceptions in the literature that use supervised learning, such as the Naïve Bayes
method or support vector machines to mine the sentiment from accounting and financial data. 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, and 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 McDonal (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 tremendously 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, mathematical proofs, and Monte Carlo simulations.

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 $1 \leq i \leq n, 1 \leq j \leq m$. 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 $A_{.,[S]}$ to denote a row vector corresponding to the $i$th 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,

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

where $y_i$ is the stock return associated with the $i$th 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 this reason and 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 a 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

\[
P(\text{sgn}(y_i) = +1) = g(p_i), \quad \text{for a monotone increasing function } g(\cdot),
\]

where \( \text{sgn}(x) \) 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:

\[
\{1, 2, \ldots, m\} = S \cup N,
\]

where \( \{1, \ldots, m\} \) is the set of indices of all words in the dictionary, \( S \) is the index set of sentiment-sensitive words, and \( N \) is the index set of sentiment-neutral words. Let \( d_{i,[S]} \) and \( d_{i,[N]} \) be the corresponding subvectors of \( d_i \), which are generated independently of each other. To model \( d_{i,[S]} \), we introduce two vectors of non-negative entries, \( O_+ \) and \( O_- \), with unit \( \ell^1 \)-norms, representing how positive and negative opinions, respectively, are distributed on the space of sentiment-sensitive words. We thereby write the sentiment-sensitive word-count generating process as:

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

where \( s_i \) is the total counts of sentiment-sensitive words in the \( i \)th article. When sentiment score \( p_i \) is high (resp. low), words in the \( i \)th article are more likely associated with \( O_+ \) (resp. \( O_- \)). Equation (4) is essentially a 2-topic topic model. Our objective is to learn these 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 it unspecified.\(^1\)

\(^1\)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. 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.
words are generated, yet is agnostic about the data generating process of the sentiment-neutral words. 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 provide details of the SSESTM procedure, which consists of three steps: the first step is to single out the set of sentiment words, \( S \); the second is to estimate the parameter, \( O \) in the topic model; the last step is to make predictions about the sentiment score of a new article based on our estimates of \( S \) and \( O \).

### 2.2 Screening Sentiment-Sensitive Words

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

We proceed using stock returns to screen out sentiment-neutral words. Intuitively, if a word appears often in articles whose tagged firms have positive returns, this word is more likely to carry information related to positive sentiment. Based on this idea, our screening step first assigns a sentiment score to each word in the dictionary, which measures how often this word is associated with an article that has a positive sentiment: for \( 1 \leq j \leq m \),

\[
f_j = \frac{\text{count of articles word } j \text{ appears with } \text{sgn}(y) = +1}{\text{count of all articles in which word } j \text{ appears}}.
\]

The next is simply to draw a line that distinguishes words with informative sentiment scores, i.e., different from 0.5. We can determine the thresholds \( (\alpha_+, \alpha_-) \), for positive and negative words, respectively, using cross-validation.

In addition to the thresholds \( (\alpha_+, \alpha_-) \), we suggest selecting a third threshold on the denominator of \( f_j \), i.e., the count of all articles in which word \( j \) appears exceeds \( \kappa \). This is because some sentiment words might happen to appear infrequently in the data sample, in which case we have very noisy information about their relevance to sentiment. Imposing this threshold is amount to taking into account the statistical accuracy of the score \( f_j \).

The definition in (5) is based on the number of articles, instead of the total number of word counts, so that a word is only counted once if it is mentioned multiple times within one article. In theory, both approaches are consistent. In practice, our approach works better because unlike nouns, sentiment words are less likely to appear multiple times in an article.

Once \( (\alpha_+, \alpha_-) \) is given, we then find all words that satisfy these restrictions and summarize them in a set, denoted as \( \hat{S} \):

\[
\hat{S} = \{ j : f_j \geq 1/2 + \alpha_+ , \text{ or } f_j \leq 1/2 - \alpha_- \} \cap \{ j : k_j \geq \kappa \},
\]

where \( k_j \) is the total count of articles in which the \( j \)th word appears.\(^2\)

\(^2\) We can also combine our vocabulary with words identified in the classical sentiment analysis, which has provided
We call this approach “screening” because for each \( j \), \( f_j \) is exactly the slope coefficient of a cross-article regression of \( \text{sgn}(y_i) \) against a dummy variable \{word \( j \) appears in article \( i \}\). Instead of \( m \) univariate regressions, an alternative approach is one multivariate regression with the perhaps sparsity assumption. We prefer the screening method, because it is simpler, and also because of the low co-occurrence of words (so that most words appear almost independently).

To justify the validity of this algorithm, we establish the so-called sure-screening property (see, e.g., Fan and Lv (2008)) of Algorithm 1 in Theorem C.2 of Appendix C. In other words, we show \( \mathbb{P}(\hat{S} = S) \) approaches one asymptotically as the number of articles, \( n \), and the number of words, \( m \), jointly increase to \( \infty \).

### 2.3 Learning Sentiment Distributions

Since we have identified the relevant wordlist \( S \), we have simplified the problem to a specific topic model with two topic vectors, summarized in a matrix \( O = [O_+, O_-] \). The matrix \( O \) determines the data generating process of the counts of sentiment-sensitive words in each article, capturing information on the frequency of words as well as their sentiment. In fact, we can rewrite these two topic vectors to a vector of frequency and a vector of tone:

\[
F = \frac{1}{2}(O_+ + O_-), \quad T = \frac{1}{2}(O_+ - O_-),
\]

(8)

where the vector of frequency characterizes the relative frequency of different words, and the vector of tone reflects the relative sentiment of different words. If a word has a larger corresponding value in \( F \), it appears more frequently. If a word has a larger value in \( T \), its sentiment is more positive. \( F \) and \( T \) are the key parameters that relate the sentiment of the \( i \)th article, \( p_i \), to the frequency and sentiment of each word therein.

In this section, we provide a supervised learning approach to estimate \( O \). Given any estimator of \( \hat{O} \), we can build estimators \( \hat{F} \) and \( \hat{T} \) 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 contain information about the sentiment of the articles, and hence can be used as training labels for a supervised learning approach. Existing topic modeling methods, (Blei et al., 2003; Anandkumar et al., 2012; Arora et al., 2013; Ke and Wang, 2017), nonetheless, are unsupervised approaches, because training labels are not available in their settings. In a low signal-to-noise ratio environment, supervised learning often leads to superior performance.

Suppose, for now, that we observe the sentiment scores of all articles in our sample. Note that various lexical vocabularies for common English writing. It is simple to expand \( \hat{S} \) to \( \tilde{S} \), by replacing \( S_2 \) in Algorithm 1:

\[
\tilde{S} = \hat{S} \cup \{1 \leq j \leq m : \max\{\ell_j, 1 - \ell_j\} \geq 1 - \beta\}.
\]

(7)

where \( \ell \in [0,1]^m \) is a vector capturing the lexical level of all dictionary words in common English writing, and \( \beta \) is a proper threshold we can tune.
model (4) implies \( \mathbb{E}\tilde{d}_i[S] := \mathbb{E}s_i^{-1}d_i[S] = p_iO_+ + (1 - p_i)O_- \). In matrix form, we have

\[
\mathbb{E}\tilde{D} = OW,
\]

where

\[
W := \begin{bmatrix}
p_1 & \cdots & p_n \\
1 - p_1 & \cdots & 1 - p_n
\end{bmatrix}, \quad \text{and} \quad \tilde{D} := [\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_n].
\]

This suggests a simple approach to recover \( O \) via a regression of \( \tilde{D} \) on \( W \).

Of course, we do not observe \( \tilde{D} \) or \( W \). It is easy to estimate \( \tilde{D} \), by plugging in \( \hat{S} \) from Algorithm 1. To estimate \( W \), we use the standardized ranks of returns as polarity scores for all articles in the training sample. More precisely, we sort the returns \( \{y_i\}_{i=1}^n \) of the training sample in the ascending order. For each \( 1 \leq i \leq n \), let

\[
\hat{p}_i = \frac{\text{rank of } y_i \text{ in all returns}}{n}.
\]

Plugging this in the definition of \( W \), yields \( \hat{W} \). This approach is very intuitive, and robust to return outliers.

We summarize the step-by-step details of our procedure in Algorithm 2. The accuracy of the estimator \( \hat{O} \) thereby hinges on the approximation quality of \( \{\hat{p}_i\}_{i=1}^n \) with respect to \( \{p_i\}_{i=1}^n \) as well as the quality of the wordlist \( \hat{S} \) obtained from screening. Theorem C.3 in Appendix C 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 will arise (unless in special cases), which depends on the correlation between the true sentiment and the estimated sentiment:

\[
\rho = \frac{12}{n} \sum_{i=1}^n \left( p_i - \frac{1}{2} \right) \left( \hat{p}_i - \frac{1}{2} \right).
\]

Specifically, the estimator \( \hat{T} \) converges to \( \rho T \) instead of \( T \). That being said, as long as the estimation quality of \( \hat{p} \) is high, \( \rho \) is fairly close to 1. It equals 1 if \( \hat{p} = p \). Moreover, what matters in practice is not the absolute magnitude of the sentiment score of each word, but the relative scores among all words. 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 \( \rho T \) 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, \( \bar{s}/|S| \leq C \) 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 \( \hat{S} \) and \( \hat{O} \). We now discuss how to estimate the polarity \( p \) for a new article with word-count \( d \in \mathbb{R}^n_+ \), which is not involved in the previous estimation.
procedure.

By (4), we have
\[ d_{[S]} \sim \text{Multinomial} \left( s, pO_+ + (1 - p)O_- \right), \]
where \( s \) is the total number of words that appear in \( S \) in this article. To estimate \( p \) using \( \hat{S} \) and \( \hat{O} \), 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.

We add a penalty \( \lambda \log(p(1-p)) \) in the likelihood function, written explicitly in (A.3) of Algorithm 3, in order to help 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 \( \lambda \). This is also equivalent to imposing a prior Beta distribution on the sentiment score. For most articles, their sentiment is neutral, so imposing such a prior improves the estimates. For articles that are not neutral, the sentiment scores will be biased towards neutrality if \( \lambda \) is large (or the prior is not uninformative). Nonetheless, the relative ranks across articles are not influenced, which convey the most useful information for the trading strategy we build in the empirical analysis.

Theorem C.4 in Appendix C 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 (\( \rho \) 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 sentiment 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.

3 Empirical Analysis

In this section, we demonstrate the empirical relevance of our text-mining framework, by directly designing trading strategies to explore the return predictability of our predicted sentiment scores.

Prior to presenting any empirical findings, it is useful to ponder over where the predictability may come 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 opens up profitable investment opportunities for high-frequency trading.

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. Therefore, we have to explore alternative data such as the news text to infer investors’ changes of expectations about firms’ future fundamentals. The signal-to-noise ratio in the bag of words we construct from news is extremely low. 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. Finally, 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 the market has fully incorporated the information therein into asset returns.

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

### 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, 1989 to July 31, 2017.

We have 22,471,222 number of unique articles after combining articles that share the same accession numbers. There are on average about 62.5% 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 16.4% articles, leaving us with a sample of 10,364,189 articles. We collect the date, the exact timestamp, tickers of the tagged firms, the title and the content of each article in this sample.

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

Note: In this table, we report the impact of each filter we apply on the number of articles in our sample. The sample period ranges from January 1, 1989 to July 31, 2017. The CRSP three-day returns are only used in training and validation steps, so we apply the CRSP filter only for articles dated from January 1, 1989 to December 31, 2012. The open-to-open returns and intraday returns are used in out-of-sample periods from February 1, 2004 to July 31, 2017.
We then match each article with the CRSP database using its tagged firm. In particular, we collect the market cap and CRSP adjusted daily close-to-close returns. We merge each article published between 4pm of day $t - 1$ and 4pm of day $t$ with its tagged firm’s three-day return (close on day $t - 2$ to close on day $t + 1$). For news that occur on holidays or weekends, we use the next available trading day as the current day $t$ and the last trading day before the news as day $t - 1$. The three-day return is more robust to the simultaneous causality problem in the sense that the news may cover some events that have occurred or some that are about to occur to its tagged firms. We only use three-day returns when training and validating the model.

We also collect open and close prices from CRSP, with which we calculate the adjusted open-to-open returns. In addition, we merge our news data with the intraday transaction prices from the NYSE Trade and Quote (TAQ) database. The open-to-open and intraday returns are only used in our out-of-sample analysis from February 2004 to July 2017. The reason we start the out-of-sample testing period from February 2004 is that since January 17, 2004 the Dow Jones Newswires data come exclusively from one datasource. Prior to then, there were multiple news sources, which may lead to redundant coverages of the same events. Although it does not affect in-sample training and validation, it could have an adverse impact on our trading strategies that rely on news being fresh. Table 1 provides details on how our data filters streamline the sample step by step.

Figure 1: Average Number of Articles per Half an Hour

![Figure 1: Average Number of Articles per Half an Hour](image)

Note: This figure plots the average numbers of articles per half an hour (24 hour EST time) from January 1, 1989 to July 31, 2017.

Figure 1 plots the average number of articles in each half hour throughout a day. The arrival of news articles is more intensive prior to US market opening and close. Figure 2 plots the average number of articles per day over a year. We observe a leap-year effect and holidays effect in the sense that non-trading days see a drop in the intensity of news articles. The rise of news articles around February, May, August, and November are due to quarterly earnings. Finally, Figure 3 plots a time
series of the total number of news articles per year in our sample. There is a steady increase in
the number of news articles until around 2007. Some patterns in the volume of news are due to
structural changes of the datasources. According to the DJN’s user guide, historically there were
three merges of news sources, which occurred on October 31, 1996, November 5, 2001, and January
16, 2004, respectively.

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.

Next, we take a few steps that are common in the field of natural language processing to clean
and structure the news articles.\textsuperscript{3} The first step is normalization, including 1) changing all words
in the article to lower case letters; 2) expanding contractions such as “haven’t” to “have not”; 3)
deleting numbers, punctuations, special symbols, and only keep English words;\textsuperscript{4} 4) combining all
paragraphs into one and eliminating spaces in between. The second step is stemming and lemmatization,
which group together the different inflected forms of a word to analyze them as a single item, e.g.,
“disappointment” to “disappoint”, “likes” to “like”, etc.\textsuperscript{5} Stemming achieves the root
form of a word even if the stem itself is not a valid English word, e.g., “accretive” and “accretion”
to “accret”. We use the most frequent word in our sample, e.g., “accretion”, among all that share
a stem to represent this stem for better interpretation. 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.\textsuperscript{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.

\textsuperscript{3}We use the natural language toolkit (NLTK) in Python to preprocess the data.
\textsuperscript{4}The list of English words is available from item 61 on http://www.nltk.org/nltk_data/.
\textsuperscript{5}The lemmatization procedure uses the WordNet as a reference database: https://wordnet.princeton.edu/.
The stemming procedure uses the package “porter2stemmer” on https://pypi.org/project/porter2stemmer/.
\textsuperscript{6}We use the list of stopwords available from item 70 on http://www.nltk.org/nltk_data/.
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. 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) lists are significantly associated with contemporaneous returns, and essentially identical in their impact.

### 3.2 Performance of Next-Day Return Predictions

To evaluate text-based trading strategies, we use models trained with a 10-year rolling sample, validated with a 5-year rolling sample, for one year ahead out-of-sample prediction and evaluation. The first training sample is from 1989 to 1998, whose validation sample is from 1999 to 2003, and the corresponding testing sample is the year of 2004. We continue to retrain our model every year, rolling all three samples forward to include one more recent year.

In each training sample, we estimate a variety of models using a grid of tuning parameters. We then use the estimated models to score each news article in the validation sample, and select the optimal set of tuning parameters which minimizes a loss function — the \( \ell^1 \)-norm of the differences between estimated scores and the standardized ranks of all articles in the validation sample.

In the out-of-sample period, we estimate the sentiment scores of articles each day, with which we

---

7The 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.

8There are four tuning parameters in our model, including \( (\alpha_+, \alpha_-, \kappa, \lambda) \). We select five choices for \( \alpha_+ \) and \( \alpha_- \) (25, 50, 100), five choices of \( \kappa \) (86%, 88%, 90%, 92%, and 94% quantiles of the count distribution each year), and three choices of \( \lambda \) (1, 5, and 10).
obtain a rank of firms that are mentioned in these articles. In the case a stock is mentioned multiple times in the news articles, we use its average score. For firms that have no news, we assign them a neutral score (i.e., 0.5). Our trading strategy is very simple. We long 50 stocks with top sentiment scores and short the bottom 50. Occasionally, there are less than 50 firms with a non-neutral score, in which case the total number of stocks we trade would be smaller than 100. Combining the long and short sides yield a zero-cost long-short portfolio.

We consider both equal-weighted and value-weighted portfolios. Value weighting imposes a strong penalty on the small caps, which might be justifiable for capacity reasons. Equal weighting is still a rather persimmons and conservative choice. At the daily frequency, the major investors are hedge funds, which are likely to use more advanced approaches than equal weighting to determine the optimal weights of different stocks.

We rebalance these portfolios only at the opening of the market on a daily basis. There are at least two reasons we choose the market opening instead of the market close. First, overnight news have not been incorporated into market prices before the opening, which is the earliest time traders can respond to information in the overnight news. Second, except for those funds that are specialized in high-frequency trading, the majority of funds are unlikely to change their positions continuously in response to intraday news, because of their capacity constraints and investment styles. Therefore, trading only at market openings allows these funds to have sufficient time to calculate their positions. Also for capacity reasons, we exclude all articles published between 9:00am and 9:30am EST, so that traders have at least 30 minutes, and usually much longer, to digest the news information and execute the trades.

*Figure 4: News Timeline*

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

*Figure 5* compares cumulative open-to-open returns of our long-short portfolios in the out-of-sample period, as well as their long and short sides, respectively. Table 2 reports the performance of these portfolios in details. We find that the equal-weighted portfolios substantially outperform their value-weighted counterparts. In particular the long-short strategy with equal weights earn an
annualized Sharpe ratio 4.50, whereas their value-weighted counterpart has 1.18. Moreover, the long positions outperform the short, with a Sharpe ratio 2.16 versus 1.33 for equally weighted portfolios. Last but not least, daily returns have significant alphas against standard benchmarks in the literature, at the cost of large turnovers. The low $R^2$s of long-short portfolios indicate that these return time series are not highly correlated with existing factors.

Figure 5: Performance Comparison of the Constructed Portfolios

![Graph showing annualized Sharpe ratios and average daily returns for different portfolios.]

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

3.3 Performance of Overnight vs Intraday Trading

We further look into the performance of these portfolios, and attribute the profits to the timing of the news articles. Specifically, our strategy trades overnight news at market opens. These news are fresh and are unlikely incorporated into the market prices yet. However, for intraday news, we build up positions only at the next-day’s market open. The time delay could be substantial. Figure 6 compares the differences of these portfolios based on the intraday and overnight news, respectively, showing that trading on overnight news indeed earns much higher returns (with a Sharpe ratio 4.06 for long-short portfolios) as opposed to trading from the next day on the intraday news (with a Sharpe ratio 1.81 for long-short portfolios). The reason that trading overnight articles earn a higher return but a slightly lower Sharpe ratio (4.06 for long-short portfolios) than trading all articles (Sharpe ratio 4.50) is that the overnight portfolio only trades part of the 100 stocks, so that its volatility is slightly higher.

The previous comparison only illustrates that the timing of the trading matters. The relatively low attribution to intraday news is mechanically due to our conservative trading strategy that delays
Table 2: Performance of the Constructed Portfolios

<table>
<thead>
<tr>
<th></th>
<th>IR</th>
<th>SR</th>
<th>Turnover</th>
<th>Annualized Return</th>
<th>FF3 alpha</th>
<th>FF3 R²</th>
<th>FF5 alpha</th>
<th>FF5 R²</th>
<th>FF5+Mom alpha</th>
<th>FF5+Mom R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>EW L-S</td>
<td>2.74</td>
<td>4.50</td>
<td>93.96%</td>
<td>0.86</td>
<td>0.87</td>
<td>0.02</td>
<td>0.85</td>
<td>0.03</td>
<td>0.85</td>
<td>0.05</td>
</tr>
<tr>
<td>EW L</td>
<td>1.96</td>
<td>2.16</td>
<td>94.48%</td>
<td>0.48</td>
<td>0.42</td>
<td>0.39</td>
<td>0.42</td>
<td>0.39</td>
<td>0.42</td>
<td>0.40</td>
</tr>
<tr>
<td>EW S</td>
<td>0.71</td>
<td>1.33</td>
<td>93.45%</td>
<td>0.38</td>
<td>0.45</td>
<td>0.33</td>
<td>0.43</td>
<td>0.34</td>
<td>0.42</td>
<td>0.36</td>
</tr>
<tr>
<td>VW L-S</td>
<td>0.48</td>
<td>1.18</td>
<td>90.44%</td>
<td>0.23</td>
<td>0.24</td>
<td>0.10</td>
<td>0.23</td>
<td>0.11</td>
<td>0.22</td>
<td>0.12</td>
</tr>
<tr>
<td>VW L</td>
<td>0.54</td>
<td>0.95</td>
<td>91.82%</td>
<td>0.20</td>
<td>0.14</td>
<td>0.30</td>
<td>0.14</td>
<td>0.30</td>
<td>0.14</td>
<td>0.30</td>
</tr>
<tr>
<td>VW S</td>
<td>-0.16</td>
<td>0.08</td>
<td>89.05%</td>
<td>0.03</td>
<td>0.10</td>
<td>0.30</td>
<td>0.09</td>
<td>0.31</td>
<td>0.09</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Note: The table reports the performance of our equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios and their long (L) and short (S) sides, respectively. The performance measures include annualized risk-adjusted alphas and $R^2$s with respect to the Fama-French three-factor model (“FF3”), the Fama-French five-factor model (“FF5”), and the Fama-French five-factor model augmented to include the momentum factor (“FF5+MOM”), as well as the daily turnover, annual information ratio, annual Sharpe ratio (SR), and unadjusted annualized returns. We define the strategy’s average daily turnover as $\text{Turnover} = \frac{1}{T} \sum_{t=1}^{T} \left( \sum_i \left| w_{i,t} \cdot (1 + r_{i,t}) - w_{i,t+1} \right| \right)$, where $w_{i,t}$ is the weight of stock $i$ in the portfolio at time $t$. The benchmark return in the calculation of the information ratio (IR) is the market ETF, SPY.

the execution of these trades to the next day’s open. We now estimate if we were to execute these trades immediately after the news events, whether this would boost the trading profits. Moreover, hypothetically, if we were to have built up the trading position for these stocks at the previous market open, how much profits we would earn if we unwind these positions at the next market open? We call this portfolio the day 0 portfolio, which is infeasible to execute, but provide a good benchmark to our feasible portfolio, which we refer to as the day+1 portfolio. For this question, we only consider the intraday news of the long-short day+1 portfolio, which only earns a 1.81 Sharpe ratio.

Figure 7 compares the cumulative returns of the day 0 and day+1 portfolios based on intraday news only, along with a sequence of portfolios that are executed at intraday time points, ranging from 3 hours before the news to 30 minutes after the news. For these hypothetical portfolios that can trade before the news, we rebalance them at the market open of the next day. For portfolios that trade only after the news, we rebalance them at the market open the day after the next day. We find that, interestingly, the earlier the trades are executed, the higher the portfolio returns. Although none of these portfolios but the day+1 portfolio are feasible in practice, they provide an interesting insight that our sentiment scores for intraday news are equally good as they are for overnight news. Trading based on these news would be equally lucrative, despite a much higher requirement on the capacity of the funds, and perhaps a much lower trading capacity in light of the transaction cost within this short period of time.

3.4 Redundant News, Prediction over Longer Horizons

We investigate for how long the return predictability persists. We consider similar portfolios as day 0 and day+1 above, but further extend it up to ten working days (two weeks) before and after the news.

---

9 Even though some of the portfolios trade the news information after the news event, we still rely on the end of the day calculations to select stocks to trade. In this regard, these portfolios are also infeasible.
Figure 6: Performance Comparison between Overnight and Intraday News

Note: This figure compares the cumulative log returns of portfolios based on different timings of the articles. The black, blue, and red lines represent the long-short (L-S), long (L), and short (S) portfolios, respectively. For the short position, the lower the curve, the higher the profits. The solid and dot-dashed lines represent overnight and intraday portfolios, respectively. All portfolios are equally weighted. The yellow solid line is the market benchmark, SPY. The legend provides a table of annual Sharpe ratios (SR) and average daily returns (AvgRet) in basis points.

Figure 7: Performance of Hypothetical Intraday Tradings

Note: This figure compares the cumulative log returns of portfolios formed based on news that occur intraday. The execution times of these portfolios are different, ranging from one day ahead of the news (day 0) portfolio, 1, 2 or 3 hours before the news, 5 minutes before the news, 15 minutes after the news, and 30 minutes after the news. The lighter the color, the earlier the execution on these news. For these hypothetical portfolios that are executed before the news, they rebalance at the market open of the next day. For the rest portfolios, they are rebalanced at the market open the day after. The day+1 portfolio executes the trades of intraday news only at the market open of the next day. None of these portfolios but the day+1 portfolio are feasible.
More specifically, for news that occur from day t-1’s open to day t’s open, we compare the cumulative returns earned separately, on day 0, (with a holding period of these stocks from day t-1’s open to day t’s open), and similarly on day±1, day±2, …, and day±10, respectively. All these portfolios are rebalanced on a daily basis. According to this definition, day+1 portfolio exactly matches our previous strategy. The day≤0 portfolios are infeasible for trading, because they require information that is only known afterwards. We include day 0 only for benchmark purpose, and day<0 to check if news are previously reported.

Figure 8 plots the cumulative returns of all these portfolios. Not surprisingly, among all feasible portfolios, the cumulative returns earned on day+1 is most significant, and the news effect quickly disappears after few days. All portfolios beyond ±2 days earn insignificant returns. This is expected, as old news get incorporated into prices and the market moves again because of the arrival of fresher news. Day-0 portfolio earns the highest return, reflecting a high contemporaneous correlation between the returns and the news sentiment. Day−1 portfolio earns an even higher return than day+1, and that day+2 portfolio earns a slightly positive return, all showing evidence of redundant news stories that have been reported few days before.

Figure 8: Performance of Day 0, Day±1, …, Day±10 Portfolios

Note: This figure compares the cumulative log returns of portfolios formed on up-to-ten working days before and after the news events. The thicker blue, red, and black lines represent the long-short portfolios of day 0, day±1, day±2, respectively. The solid lines are feasible portfolios because they are formed after the occurrence of news. The dot-dashed lines are benchmark portfolios despite being infeasible to execute. We use thinner dotted lines with various colors to denote the remaining portfolios that earn insignificant returns.

3.5 Comparison with the LM Dictionary and the RavenPack Approach

We compare SSESTM with the LM 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 their Fin-Neg list is more relevant because words therein typically have negative implications in a financial sense.

We also compare SSESTM 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 February 2004. The relevance score is between 0 and 100, indicating how strongly related the tagged firm is to the underlying news story. According to their user guide, 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.

To build long and short strategies using these alternatives, for LM we sort stocks based on the tf-idf weighted counts of positive words minus negative words, as they recommend, and long 50 stocks that rank on top and short 50 stocks that rank at the bottom. For RP, we use a similar strategy to ours by sorting on RavenPack’s CSS. Figure 9 compares the cumulative returns of all long-short strategies. Our SSESTM earns substantially higher Sharpe ratios than these two alternatives, regardless of the portfolio weighting schemes. The gain in performance against LM comes from the differences in our scoring strategies. Our choice of wordlist and the weights of words are entirely data-driven, whereas their wordlist is more subjective, and their weights of words are adhoc. RavenPack’s black-box strategy is a bit better than LM, but still dominated by our SSESTM, which is a white-box with theoretical guarantees.

### 3.6 Transaction Cost

The performance analysis of the above portfolios ignore the transaction costs. After all, the primary purpose of this paper is to provide sentiment scores for news articles and show that they can predict stock returns at short-term horizons, rather than to construct profitable trading strategies in practice. That said, it does not hurt to show empirically that we can reduce the portfolio turnovers and obtain a viable trading strategy which takes into account some conservative estimates of the transaction cost, i.e., transaction cost = 2× turnover × 10 bps.

To lower the turnover of the portfolios, we can simply use the smoothed sentiment scores to sort stocks. We adopt the widely used exponential smoothing scheme, which constructs a new sequence of scores based on a history of raw estimates using a recursive formula: \( \tilde{s}_t = \iota \tilde{s}_{t-1} + (1-\iota)\tilde{s}_{t-1} \). This simple strategy would work if the reduction in turnover (and hence the transaction cost) overshadows the deterioration of the performance of the portfolio.

Table 3 reports the performance of SSESTM portfolios based on sentiment scores with various

---

degrees of smoothness. As the smoothing parameter $\iota$ decrease, the turnover drops accordingly, and not surprisingly, the performance of the corresponding portfolios also deteriorate a bit. The sweet point appears to be around $\alpha = 0.6$.

### 3.7 To What Words Returns are Most Sensitive?

Finally we investigate the list of words that tend to be associated with the sign of returns selected using the training and validation samples. Because there are moderate variations in the selected words across different training periods, we report the top 50 selected words for positive and negative sentiment, respectively.

Among all positive words we tease out, “undervaluing”, “repurchase”, “surpass”, “surge”, and “beat” consistently rank among the most optimistic words; some warfare related words surprisingly show up moderately frequent: “naval”, “warfare”, “troop”, “armor”, “maritime”, “tanker”, “trinitrotoluene”; there are also words that appear often but are not clearly associated with a positive sentiment, such as “fossil”, “chest”, “tiffany”, which might be the noise.

The negative words, on the other hand, clearly demonstrate pessimistic sentiment. Several words are persistently selected in all periods, including “shortfall”, “downgrade”, “disappointing”, “tumble”, “blame”, “hurt”, “plummet”, “plunge”, “auditor”, and “slovenliness.” The rest of the words are also quite self-explanatory. Very few words appear less meaningful, such as “bulletin”, “subplot”, “covenant”, “qualification”, and “banana”.

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
Table 3: Performance of the Long-Short Portfolios Net Transaction Costs

<table>
<thead>
<tr>
<th>τ</th>
<th>Average Return Before Cost</th>
<th>Turnover</th>
<th>Sharpe Ratio</th>
<th>Average Return Net Cost</th>
<th>Sharpe Ratio Net Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>86.21%</td>
<td>91.94%</td>
<td>4.54</td>
<td>39.87%</td>
<td>2.07</td>
</tr>
<tr>
<td>0.8</td>
<td>86.62%</td>
<td>89.62%</td>
<td>4.57</td>
<td>41.45%</td>
<td>2.16</td>
</tr>
<tr>
<td>0.7</td>
<td>85.23%</td>
<td>86.98%</td>
<td>4.42</td>
<td>41.39%</td>
<td>2.11</td>
</tr>
<tr>
<td>0.6</td>
<td>84.86%</td>
<td>84.19%</td>
<td>4.39</td>
<td>42.43%</td>
<td>2.17</td>
</tr>
<tr>
<td>0.5</td>
<td>83.16%</td>
<td>81.38%</td>
<td>4.28</td>
<td>42.14%</td>
<td>2.14</td>
</tr>
<tr>
<td>0.4</td>
<td>80.00%</td>
<td>78.47%</td>
<td>4.07</td>
<td>40.45%</td>
<td>2.03</td>
</tr>
<tr>
<td>0.3</td>
<td>73.56%</td>
<td>75.30%</td>
<td>3.67</td>
<td>35.61%</td>
<td>1.75</td>
</tr>
<tr>
<td>0.2</td>
<td>68.18%</td>
<td>71.69%</td>
<td>3.46</td>
<td>32.05%</td>
<td>1.60</td>
</tr>
<tr>
<td>0.15</td>
<td>61.89%</td>
<td>69.94%</td>
<td>3.11</td>
<td>26.79%</td>
<td>1.31</td>
</tr>
<tr>
<td>0.1</td>
<td>52.43%</td>
<td>67.35%</td>
<td>2.47</td>
<td>18.48%</td>
<td>0.83</td>
</tr>
<tr>
<td>0.05</td>
<td>33.34%</td>
<td>64.69%</td>
<td>1.45</td>
<td>0.74%</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

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

clarity are highly desirable.

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. Our SSESTM approach delivers superior out of sample performance than the alternatives based on a subjective dictionary or a commercial black-box by a large margin. We find evidence that prices respond to market news within minutes. The sooner the portfolios are built, the more the profits, despite that the predictability disappears after one day or two.
### Table 4: The List of Words via Screening

<table>
<thead>
<tr>
<th>Positive</th>
<th>f_j</th>
<th>t_j</th>
<th>k_j</th>
<th>Negative</th>
<th>f_j</th>
<th>t_j</th>
<th>k_j</th>
</tr>
</thead>
<tbody>
<tr>
<td>undervalue</td>
<td>0.605</td>
<td>14</td>
<td>2533</td>
<td>shortfall</td>
<td>0.323</td>
<td>14</td>
<td>8357</td>
</tr>
<tr>
<td>repurchase</td>
<td>0.572</td>
<td>14</td>
<td>48247</td>
<td>downgrade</td>
<td>0.381</td>
<td>14</td>
<td>21946</td>
</tr>
<tr>
<td>surpass</td>
<td>0.556</td>
<td>14</td>
<td>6624</td>
<td>disappointing</td>
<td>0.392</td>
<td>14</td>
<td>19530</td>
</tr>
<tr>
<td>surge</td>
<td>0.551</td>
<td>13</td>
<td>10891</td>
<td>tumble</td>
<td>0.403</td>
<td>14</td>
<td>3801</td>
</tr>
<tr>
<td>fossil</td>
<td>0.564</td>
<td>12</td>
<td>2086</td>
<td>blame</td>
<td>0.413</td>
<td>14</td>
<td>6761</td>
</tr>
<tr>
<td>naval</td>
<td>0.551</td>
<td>12</td>
<td>3321</td>
<td>hurt</td>
<td>0.415</td>
<td>14</td>
<td>27726</td>
</tr>
<tr>
<td>beat</td>
<td>0.557</td>
<td>11</td>
<td>22470</td>
<td>plummet</td>
<td>0.417</td>
<td>14</td>
<td>2197</td>
</tr>
<tr>
<td>benign</td>
<td>0.552</td>
<td>11</td>
<td>2096</td>
<td>plunge</td>
<td>0.422</td>
<td>14</td>
<td>5285</td>
</tr>
<tr>
<td>jump</td>
<td>0.550</td>
<td>11</td>
<td>21370</td>
<td>auditor</td>
<td>0.427</td>
<td>14</td>
<td>8650</td>
</tr>
<tr>
<td>warfare</td>
<td>0.560</td>
<td>10</td>
<td>2060</td>
<td>slovenliness</td>
<td>0.433</td>
<td>14</td>
<td>16234</td>
</tr>
<tr>
<td>upward</td>
<td>0.554</td>
<td>10</td>
<td>5566</td>
<td>bulletin</td>
<td>0.409</td>
<td>13</td>
<td>3394</td>
</tr>
<tr>
<td>chest</td>
<td>0.553</td>
<td>10</td>
<td>1364</td>
<td>sluggish</td>
<td>0.431</td>
<td>13</td>
<td>5067</td>
</tr>
<tr>
<td>imbalance</td>
<td>0.565</td>
<td>9</td>
<td>2976</td>
<td>woe</td>
<td>0.434</td>
<td>12</td>
<td>2271</td>
</tr>
<tr>
<td>unsolicited</td>
<td>0.556</td>
<td>9</td>
<td>2852</td>
<td>downward</td>
<td>0.434</td>
<td>12</td>
<td>6294</td>
</tr>
<tr>
<td>customarily</td>
<td>0.552</td>
<td>8</td>
<td>11033</td>
<td>warn</td>
<td>0.438</td>
<td>12</td>
<td>22535</td>
</tr>
<tr>
<td>depositary</td>
<td>0.549</td>
<td>8</td>
<td>39034</td>
<td>halt</td>
<td>0.417</td>
<td>11</td>
<td>8354</td>
</tr>
<tr>
<td>tiffany</td>
<td>0.548</td>
<td>8</td>
<td>2104</td>
<td>waiver</td>
<td>0.422</td>
<td>11</td>
<td>4159</td>
</tr>
<tr>
<td>troop</td>
<td>0.544</td>
<td>8</td>
<td>2193</td>
<td>lower</td>
<td>0.425</td>
<td>11</td>
<td>52652</td>
</tr>
<tr>
<td>armor</td>
<td>0.544</td>
<td>8</td>
<td>3062</td>
<td>subplot</td>
<td>0.428</td>
<td>11</td>
<td>1974</td>
</tr>
<tr>
<td>bought</td>
<td>0.554</td>
<td>7</td>
<td>2470</td>
<td>resign</td>
<td>0.442</td>
<td>11</td>
<td>25239</td>
</tr>
<tr>
<td>maritime</td>
<td>0.554</td>
<td>7</td>
<td>2966</td>
<td>qualification</td>
<td>0.443</td>
<td>11</td>
<td>16307</td>
</tr>
<tr>
<td>tanked</td>
<td>0.552</td>
<td>7</td>
<td>2687</td>
<td>postpone</td>
<td>0.444</td>
<td>11</td>
<td>5496</td>
</tr>
<tr>
<td>deepwater</td>
<td>0.545</td>
<td>7</td>
<td>4051</td>
<td>fell</td>
<td>0.431</td>
<td>10</td>
<td>68610</td>
</tr>
<tr>
<td>bend</td>
<td>0.552</td>
<td>6</td>
<td>1592</td>
<td>covenant</td>
<td>0.432</td>
<td>10</td>
<td>10093</td>
</tr>
<tr>
<td>climb</td>
<td>0.551</td>
<td>6</td>
<td>11324</td>
<td>delay</td>
<td>0.433</td>
<td>10</td>
<td>39783</td>
</tr>
<tr>
<td>declare</td>
<td>0.549</td>
<td>6</td>
<td>42580</td>
<td>lackluster</td>
<td>0.438</td>
<td>10</td>
<td>2573</td>
</tr>
<tr>
<td>unanimously</td>
<td>0.544</td>
<td>6</td>
<td>5730</td>
<td>inefficiency</td>
<td>0.440</td>
<td>10</td>
<td>3829</td>
</tr>
<tr>
<td>pentagon</td>
<td>0.543</td>
<td>6</td>
<td>2401</td>
<td>issuable</td>
<td>0.441</td>
<td>10</td>
<td>4006</td>
</tr>
<tr>
<td>pep</td>
<td>0.543</td>
<td>6</td>
<td>2957</td>
<td>soften</td>
<td>0.443</td>
<td>10</td>
<td>5236</td>
</tr>
<tr>
<td>hart</td>
<td>0.541</td>
<td>6</td>
<td>5376</td>
<td>hamper</td>
<td>0.446</td>
<td>10</td>
<td>3162</td>
</tr>
<tr>
<td>terrain</td>
<td>0.540</td>
<td>6</td>
<td>2583</td>
<td>miss</td>
<td>0.447</td>
<td>10</td>
<td>23521</td>
</tr>
<tr>
<td>scap</td>
<td>0.584</td>
<td>5</td>
<td>1209</td>
<td>banana</td>
<td>0.426</td>
<td>9</td>
<td>1019</td>
</tr>
<tr>
<td>buoy</td>
<td>0.581</td>
<td>5</td>
<td>625</td>
<td>slash</td>
<td>0.435</td>
<td>9</td>
<td>6617</td>
</tr>
<tr>
<td>oversold</td>
<td>0.568</td>
<td>5</td>
<td>1084</td>
<td>insufficient</td>
<td>0.437</td>
<td>9</td>
<td>2450</td>
</tr>
<tr>
<td>scudder</td>
<td>0.564</td>
<td>5</td>
<td>2545</td>
<td>notwithstanding</td>
<td>0.438</td>
<td>9</td>
<td>1947</td>
</tr>
<tr>
<td>midstream</td>
<td>0.546</td>
<td>5</td>
<td>3648</td>
<td>default</td>
<td>0.440</td>
<td>9</td>
<td>8215</td>
</tr>
<tr>
<td>liquefy</td>
<td>0.546</td>
<td>5</td>
<td>2415</td>
<td>soft</td>
<td>0.440</td>
<td>9</td>
<td>19819</td>
</tr>
<tr>
<td>upgrade</td>
<td>0.544</td>
<td>5</td>
<td>37417</td>
<td>irregularity</td>
<td>0.442</td>
<td>9</td>
<td>2296</td>
</tr>
<tr>
<td>turkey</td>
<td>0.541</td>
<td>5</td>
<td>5017</td>
<td>worse</td>
<td>0.444</td>
<td>9</td>
<td>2501</td>
</tr>
<tr>
<td>tender</td>
<td>0.541</td>
<td>5</td>
<td>21632</td>
<td>unlawful</td>
<td>0.447</td>
<td>9</td>
<td>13727</td>
</tr>
<tr>
<td>crane</td>
<td>0.540</td>
<td>5</td>
<td>3618</td>
<td>unfortunately</td>
<td>0.448</td>
<td>9</td>
<td>5063</td>
</tr>
<tr>
<td>bracket</td>
<td>0.540</td>
<td>5</td>
<td>2897</td>
<td>reassess</td>
<td>0.427</td>
<td>8</td>
<td>1404</td>
</tr>
<tr>
<td>propane</td>
<td>0.539</td>
<td>5</td>
<td>3910</td>
<td>defibrinate</td>
<td>0.441</td>
<td>8</td>
<td>28366</td>
</tr>
<tr>
<td>programma</td>
<td>0.538</td>
<td>5</td>
<td>2670</td>
<td>erosion</td>
<td>0.441</td>
<td>8</td>
<td>2722</td>
</tr>
<tr>
<td>terrific</td>
<td>0.571</td>
<td>4</td>
<td>1819</td>
<td>weak</td>
<td>0.445</td>
<td>8</td>
<td>51938</td>
</tr>
<tr>
<td>clip</td>
<td>0.565</td>
<td>4</td>
<td>1031</td>
<td>deficiency</td>
<td>0.449</td>
<td>8</td>
<td>6386</td>
</tr>
<tr>
<td>trinitrotoluene</td>
<td>0.564</td>
<td>4</td>
<td>2002</td>
<td>withhold</td>
<td>0.449</td>
<td>8</td>
<td>2100</td>
</tr>
<tr>
<td>potent</td>
<td>0.563</td>
<td>4</td>
<td>1344</td>
<td>erode</td>
<td>0.454</td>
<td>8</td>
<td>3134</td>
</tr>
<tr>
<td>milk</td>
<td>0.561</td>
<td>4</td>
<td>777</td>
<td>unfavorable</td>
<td>0.436</td>
<td>7</td>
<td>5929</td>
</tr>
<tr>
<td>tout</td>
<td>0.560</td>
<td>4</td>
<td>1127</td>
<td>widen</td>
<td>0.437</td>
<td>7</td>
<td>10503</td>
</tr>
</tbody>
</table>

Note: In this table, we provide top 50 lists of words with positive and negative sentiment, respectively, based on our screening using the 14 separate training and validation samples. These 50 words are selected by first sorting on the number of appearance, t_j, in the 14 screening outcomes, and then sorting on their average sentiment scores f_j. We also report the average number of articles, k_j, in which each word appear in sample.
References


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


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


Appendix

A Algorithms

Algorithm 1.

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

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

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

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

where $k_j$ is the total count of articles in which the $j$th word appears.

Algorithm 2.

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

$$\hat{p}_i = \text{rank of } y_i \text{ in all returns } n.$$

(A.1)

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

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

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

Algorithm 3.

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

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

where $d_j$, $\hat{O}_{+,j}$, and $\hat{O}_{-,j}$ are the $j$th entries of the corresponding vectors, and $\lambda > 0$ is a tuning parameter.
B Monte Carlo Simulations

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

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

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

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

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

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

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

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

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

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

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

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

Figure A.2: Estimation Results in Simulations

Note: This figure compares the averages of $\hat{F}$ (blue, solid) and $\hat{T}$ (red, solid) across Monte Carlo repetitions with $F$ (black, dotted), $T$ (thin, black, dashed), and $\rho T$ (thick, black, dashed), respectively, using the benchmark parameters. The blue and red dotted lines plot the 2.5% and 97.5% quantiles of the Monte Carlo estimates.
estimated scores towards 0.5, although it has no impact on their ranks. In the benchmark setting, the average correlation across all Monte Carlo repetitions is 0.85 with a standard deviation 0.0014. If we decrease \( \bar{s} \) from 10 to 5, the quality of the estimates becomes worse due to less observations from words in \( S \). Similarly, when decrease \( n \) to 5,000, the estimates become less accurate, since the sample size is smaller. If the size of the dictionary, \( m \), or the size of the dictionary of the sentiment words, \(|S|\), drop by half, the estimates improve, despite that the improvement is marginal. Overall, these observations match what the statistical theory predicts.

<table>
<thead>
<tr>
<th>Table A.1: Spearman’s Correlation Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Avg S-Corr} )</td>
</tr>
<tr>
<td>---------------------------</td>
</tr>
<tr>
<td>Avg S-Corr</td>
</tr>
<tr>
<td>Std Dev</td>
</tr>
</tbody>
</table>

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

## C Statistical Theory

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

\[
\hat{S} = \{ j : f_j \geq 1/2 + \alpha_+ \} \cup \{ j : f_j \leq 1/2 - \alpha_- \}.
\]

where

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

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

### C.1 Regularity Conditions

Let \( s_{\text{max}}, s_{\text{min}}, \) 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 \( \Omega_i = \mathbb{E}d_{i,[N]} \).\(^{11}\) For each \( j \in N \), let \( \Omega_{\text{min},j}, \Omega_{\text{max},j}, \) and \( \Omega_{\cdot,j} \) be the maximum, minimum, and average of \( \{\Omega_{i,j}\}_{i=1}^n \), respectively.

\(^{11}\)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 \( q_i \in \mathbb{R}^{\mid N\mid} \) is a distribution on the space of \( N \), then \( \Omega_i = n_i q_i \).
We assume
\[
\frac{s_{\text{max}}}{s} \leq C, \quad \max_{j \in \mathcal{N}} \Omega_{\text{max},j} = C, \quad \min_{j \in \mathcal{S}} n\tilde{s}(O_{+j} + O_{-j}) \rightarrow \infty, \quad \min_{j \in \mathcal{N}} \frac{n\tilde{\Omega}_{-j}}{\log(m)} \rightarrow \infty. \quad (C.6)
\]

The last two inequalities in (C.6) require the expected count of any word in all of \( n \) training articles to be much larger than \( \log(m) \). Since \( n \) is large in real data, this condition is mild. For a constant \( c_0 \in (0, 1) \), we assume
\[
\min_{j \in \mathcal{S}} \sum_{i=1}^{n} s_i \left[ p_i O_{+j} + (1 - p_i) O_{-j} \right] \geq c_0. \quad (C.7)
\]
This condition says that the expected count of a word \( j \in \mathcal{S} \) in all training articles cannot be much smaller than \( n\tilde{s}F_j \), where \( F_j \) is the vector of frequency defined in (8). It is for technical convenience.

We also assume
\[
\frac{1}{n} \sum_{i=1}^{n} p_i = \frac{1}{2}, \quad \frac{1}{n} \sum_{i=1}^{n} s_i \mathbb{E}[\text{sgn}(y_i)] = 0, \quad (C.8)
\]
This condition essentially requires that we have approximately equal number of articles with positive and negative tone. Note that we can always keep the same number of articles associated with positive and negative returns in the training stage, so this condition is mild. We also assume
\[
\frac{1}{n} \sum_{i=1}^{n} \Omega_{i,j} \mathbb{E}[\text{sgn}(y_i)] = 0, \quad \text{for all } j \in \mathcal{N}. \quad (C.9)
\]
This condition ensures that the count of any sentiment-neutral word has no correlation with the sign of the stock returns (so they are indeed “sentiment-neutral”). All equalities in (C.8)-(C.9) do not need to hold exactly. We impose exact equalities so that the conclusions are more elegant.

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

First, we consider the screening step. We define a quantity to capture the sensitivity of stock returns to article polarity:
\[
\theta \equiv \frac{\sum_{i=1}^{n} s_i \left( p_i - \frac{1}{2} \right) \left[ g(p_i) - \frac{1}{2} \right]}{\sum_{i=1}^{n} s_i}, \quad (C.10)
\]
where \( g(\cdot) \) is the monotone increasing function defined in (2). When \( g(\frac{1}{2}) = \frac{1}{2} \), this quantity is lower bounded by \( \frac{1}{n^2} \sum_{i=1}^{n} s_i (p_i - \frac{1}{2})^2 \). Roughly speaking, \( \theta \) measures the steepness of \( g \) and the extremeness of training articles’ polarities.

**Theorem C.1.** Consider the model (1)-(4), where (C.6)-(C.9) hold. As \( n, m \rightarrow \infty \), with probability \( 1 - o(1) \),
\[
|f_j - 1/2| \begin{cases} \geq 2 \theta \frac{\left| O_{+j} - O_{-j} \right|}{\Omega_{+j} + \Omega_{-j}} + \frac{C \sqrt{\log(m)}}{\sqrt{n \min\{1, \tilde{s}(O_{+j} + O_{-j})\}}} \quad \text{for } j \in \mathcal{S}, \\ \leq \frac{C \sqrt{\log(m)}}{\sqrt{n \min\{1, \Omega_{-j}\}}} \quad \text{for } j \in \mathcal{N}. \end{cases}
\]

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

\textbf{Theorem C.2} (Sure Screening). \textit{Consider the model (1)-(4), where (C.6)-(C.9) hold. We assume
\( n \theta^2 \min_{j \in S} \frac{(O_{+j} - O_{-j})^2}{(O_{+j} + O_{-j})^2} \geq \frac{\log^2(m)}{\min\{1, \bar{s} \min_{j \in S} (O_{+j} + O_{-j}), \min_{j \in N} \bar{O}_{-j}\}} \).\textsuperscript{(C.11)}

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

The desired number of training articles for sure screening is determined by three factors. First, \( \theta \). The sensitivity of stock returns to article polarity, defined in (C.10). Second, \( \min_{j \in S} \frac{|O_{+j} - O_{-j}|}{O_{+j} + O_{-j}} \). It represents the word’s frequency-adjusted sentiment. Third, \( \min\{1, \bar{s} \min_{j \in S} (O_{+j} + O_{-j}), \min_{j \in N} \bar{O}_{-j}\} \). 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 \( O_- \).

\textbf{Theorem C.3} (Estimation Error of Sentiment Vectors). \textit{Consider the model (1)-(4), where (C.6)-(C.9) and (C.11) hold. With probability \( 1 - o(1) \),
\[ \| \hat{F} - F \|_1 \leq C \sqrt{\frac{|S| \log(m)}{n \bar{s}}} \],
\[ \| \hat{T} - \rho T \|_1 \leq C \sqrt{\frac{|S| \log(m)}{n \bar{s}}} \].

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 \( \ell^1 \)-norm between true and estimated topic vectors is
\[ \sqrt{\frac{|S|}{n \bar{s}}} \], up to a logarithmic factor.

Our model imposes a 2-topic topic model on sentiment-sensitive words, so the intrinsic dictionary 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
Given a new article with polarity \( p \), define the \textit{rescaled polarity} as

\[
p^* = \frac{1}{2} + \rho^{-1}(p - \frac{1}{2}).
\]  

(C.12)

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

\textbf{Theorem C.4} (Scoring Error on New Article). Consider the model (1)-(4), where (C.6)-(C.9) hold. Define \( O^{(\rho)} = [O^{(\rho)}_+, O^{(\rho)}_-] \), with \( O^{(\rho)}_\pm = F \pm \rho T \). Suppose (C.11) is satisfied with \( O \) replaced by \( O^{(\rho)} \). Let \( d \in \mathbb{R}^m_+ \) be the word count vector of a new article with 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

\[
\text{err}_n = \frac{1}{p\sqrt{\Theta}} \left( \frac{\sqrt{|S|\log(m)}}{\rho \sqrt{ns\Theta}} + \frac{1}{\sqrt{s}} \right), \text{ where } \Theta = \sum_{j \in S} \frac{(O_{+,j} - O_{-,j})^2}{O_{+,j} + O_{-,j}}.
\]

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

\[
|\hat{p} - p^*| \leq C \min\left\{ 1, \frac{\rho^2\Theta}{\lambda} \right\} \text{err}_n + C \min\left\{ 1, \frac{\lambda}{\rho^2\Theta} \right\} |p^* - \frac{1}{2}|.
\]

Therefore, the optimal choice of tuning parameter is \( \lambda = \frac{\rho^2\Theta}{|p^* - \frac{1}{2}|} \text{err}_n \), and the associated scoring error is \( |\hat{p} - p^*| \leq C \min\{\text{err}_n, |p^* - \frac{1}{2}|\} \).

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 \( |\hat{p}_i - p_i| \) is small in the training stage, then \( p^* \approx p \), and our method also correctly estimates \( p \).

The choice of \( \lambda \) yields a bias-variance trade-off. In the error bound for \( |\hat{p} - p^*| \), the first term \( \min\{1, \frac{\rho^2\Theta}{\lambda} \} \text{err}_n \) is the “variance” term, decreasing with \( \lambda \); the second term \( \min\{1, \frac{\lambda}{\rho^2\Theta} \} |p^* - \frac{1}{2}| \) is the “bias” term, increasing with \( \lambda \). 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

\[12\] By the way of construction, \( \hat{p}_i \) is uniformly distribution on \([0, 1]\). So, when \( p_i = \hat{p}_i \), we have \( \rho = 12 \int_0^1 (x - \frac{1}{2})^2 dx = 1.\]
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 (C.12), which means in some scenarios our estimator does a better job estimating the original \( p \).

The error rate \( \text{err}_n \) 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 \( \text{err}_n \to 0 \), we need that the length of the new article goes to infinity asymptotically. Nonetheless, the length of training articles can be finite.

D Mathematical Proofs

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

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

\[
f_j = \frac{1}{2} \left( 1 + \frac{L^+_j - L^-_j}{2L^+_j + L^-_j} \right) = \frac{1}{2} + \frac{\sum_{i=1}^n t_i \cdot d_{i,j}}{\sum_{i=1}^n d_{i,j}}.
\]

(D.13)

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

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

\[d_{i,j} \sim \text{Binomial}(s_i, F_j + \eta_iT_j).\]

(D.14)

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

\[
f_j \equiv \frac{1}{2} + \frac{\sum_{i=1}^n \sum_{\ell=1}^{s_i} t_i \cdot b_{i,j,\ell}}{\sum_{i=1}^n \sum_{\ell=1}^{s_i} b_{i,j,\ell}}, \quad \text{where} \quad b_{i,j,\ell} \overset{\text{iid}}{\sim} \text{Bernoulli}(F_j + \eta_iT_j).
\]

(D.15)

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

\[
\left| \sum_{i=1}^n \sum_{\ell=1}^{s_i} b_{i,j,\ell} - \sum_{i=1}^n s_i(F_j + \eta_iT_j) \right| \leq C \sqrt{\sum_{i=1}^n 2s_iF_j \log(m) + \log(m)}
\leq C \sqrt{n\bar{s}F_j \log(m) + \log(m)}
\leq C \sqrt{n\bar{s}F_j \log(m)},
\]

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

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

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

\[
f_j = \frac{1}{2} + \frac{\sum_{i=1}^n t_i s_i (F_j + \eta_i T_j) + O(\sqrt{n\bar{s}F_j \log(m)})}{\sum_{i=1}^n s_i (F_j + \eta_i T_j) + O(\sqrt{n\bar{s}F_j \log(m)})}
= \frac{1}{2} + \frac{F_j \sum_{i=1}^n t_i s_i + T_j \sum_{i=1}^n t_i \eta_i s_i + O(\sqrt{n\bar{s}F_j \log(m)})}{F_j \sum_{i=1}^n s_i + T_j \sum_{i=1}^n \eta_i s_i + O(\sqrt{n\bar{s}F_j \log(m)})}.
\]

(D.16)

In the denominator, the sum of the first two terms can be rewritten as \( \sum_{i=1}^n s_i [p_iO_{+j} + (1 - p_i)O_{-j}] \). It is upper bounded by \( 2n\bar{s}F_j \), and by (C.7), it is also lower bounded by \( 2c_0n\bar{s}F_j \). Furthermore, since \( n\bar{s}F_j \gg \log(m) \), the last term is negligible compared to the first two terms. Hence, the denominator in (D.16) is between \( c_0n\bar{s}F_j \) and \( 4n\bar{s}F_j \). It follows that

\[
|f_j - 1/2| \geq \frac{|T_j \sum_{i=1}^n t_i \eta_i s_i|}{4n\bar{s}F_j} - \frac{|F_j \sum_{i=1}^n t_i s_i|}{c_0n\bar{s}F_j} + O\left(\frac{\sqrt{n\bar{s}F_j \log(m)}}{c_0n\bar{s}F_j}\right).
\]

(D.17)

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

with probability \( 1 - O(m^{-2}) \),

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

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

\[
|f_j - 1/2| \geq \frac{|T_j |2n\bar{s}\theta|}{4n\bar{s}F_j} - \frac{F_j \cdot C\bar{s} \sqrt{n \log(m)}}{c_0n\bar{s}F_j} + O\left(\frac{\sqrt{n\bar{s}F_j \log(m)}}{c_0n\bar{s}F_j}\right)
\geq \frac{\theta |T_j|}{2F_j} + O\left(\frac{\log(m)}{n}\right) + O\left(\frac{\log(m)}{n\bar{s}F_j}\right).
\]

(D.18)

This gives the first claim of Theorem C.1.

Consider \( j \in N \). We model that \( d_{i,[N]} \) follows a multinomial distribution with \( \mathbb{E}d_{i,[N]} = \Omega_i \). Equivalently, \( d_{i,[N]} \sim \text{Multinomial}(k_i, q_i) \), where \( k_i \) is the count of all words from \( N \) in article \( i \) and \( q_i \equiv \Omega_i \). 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, \( \sum_{i=1}^{n} k_{i}q_{i,j} = \sum_{i=1}^{n} \Omega_{i,j} = n\bar{\Omega}_{i,j} \),
where by (C.6), \( n\bar{\Omega}_{i,j} \gg \log(m) \). Therefore, we have
\[
\left| \sum_{i=1}^{n} d_{i,j} - \sum_{i=1}^{n} \Omega_{i,j} \right| \leq C \sqrt{n\bar{\Omega}_{i,j} \log(m)}.
\]
Similarly, conditioning on \( \{ t_{i} \}_{i=1}^{n} \), with probability \( 1 - O(m^{-2}) \),
\[
\left| \sum_{i=1}^{n} t_{i}d_{i,j} - \sum_{i=1}^{n} t_{i}\Omega_{i,j} \right| \leq C \sqrt{n\bar{\Omega}_{i,j} \log(m)}.
\]
Plugging them into (D.13) gives
\[
f_{j} = \frac{1}{2} + \frac{\sum_{i=1}^{n} t_{i}\Omega_{i,j} + O([n\bar{\Omega}_{i,j} \log(m)]^{1})}{\sum_{i=1}^{n} \Omega_{i,j} + O([n\bar{\Omega}_{i,j} \log(m)]^{1})}
= \frac{1}{2} + \frac{\sum_{i=1}^{n} t_{i}\Omega_{i,j} + O([n\bar{\Omega}_{i,j} \log(m)]^{1})}{n\bar{\Omega}_{i,j} + O([n\bar{\Omega}_{i,j} \log(m)]^{1})}.
\]
We then deal with the randomness of \( \{ t_{i} \}_{i=1}^{n} \). By Hoeffding’s inequality, with probability \( 1 - O(m^{-2}) \),
\[
| \sum_{i=1}^{n} \Omega_{i,j}(t_{i} - \mathbb{E}t_{i}) | \leq C \sqrt{\sum_{i=1}^{n} \Omega_{i,j}^{2} \log(m)} \leq C\bar{\Omega}_{i,j} \sqrt{n \log(m)},
\]
where the last inequality is from the condition \( \Omega_{\max,j} \leq C\bar{\Omega}_{i,j} \). Moreover, by our condition (C.8), \( \sum_{i=1}^{n} \Omega_{i,j} \mathbb{E}t_{i} = 0 \). The above imply
\[
\left| \sum_{i=1}^{n} t_{i}\Omega_{i,j} \right| \leq C\bar{\Omega}_{i,j} \sqrt{n \log(m)}.
\]
We plug it into (D.19) and note that the denominator of (D.19) is \( \geq n\bar{\Omega}_{i,j} \), since \( n\bar{\Omega}_{i,j} \gg \log(m) \). It follows that
\[
| f_{j} - 1/2 | \leq \frac{C\bar{\Omega}_{i,j} \sqrt{n \log(m)} + O([n\bar{\Omega}_{i,j} \log(m)]^{1})}{n\bar{\Omega}_{i,j}}
\leq O\left( \frac{\sqrt{\log(m)}}{n} \right) + O\left( \frac{\log(m)}{n^{2} \bar{\Omega}_{i,j}} \right).
\]
This gives the second claim of Theorem C.1.

Next, we prove Theorem C.2. By (D.18) and (D.20), with probability \( 1 - O(m^{-1}) \), simultaneously for all \( 1 \leq j \leq m \),
\[
| f_{j} - 1/2 | \begin{cases} 
\geq \frac{\theta_{T_{j}}}{2F_{j}} + O(e_{n}), & j \in S, \\
\leq O(e_{n}), & j \in N,
\end{cases}
\]
where \( e_{n}^{2} = (\min\{ 1, \bar{s} \min_{j \in S} F_{j}, \min_{j \in N} \bar{\Omega}_{i,j} \}^{-1} \log(m))^{1/2} \). The assumption (C.11) ensures that \( \theta_{T_{j}} \gg e_{n} \sqrt{\log(m)} \). By setting the threshold at \( e_{n} \sqrt{\log(\log(m)))} \), all words in \( S \) will retain and all words in \( N \) will be screened out. □
D.2 Proof of Theorem C.3

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

\[
O^* = O\hat{W}(\hat{W}\hat{W}')^{-1}.
\]

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

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

In the second part of our proof, we show that

\[
\|\hat{O}_\pm - O^*_\pm\|_1 \leq C \sqrt{|S| \log(m)/(ns)} \tag{D.22}
\]

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

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

\[
[F^*, T^*] = O^* \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} = O(\hat{W}\hat{W}')(\hat{W}\hat{W}')^{-1} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} = [F, T] S \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} (\hat{W}\hat{W}')(\hat{W}\hat{W}')^{-1} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} = M. \tag{D.23}
\]

We now calculate the \( 2 \times 2 \) matrix \( M \). With the returns sorted in the ascending order, \( y_1 < y_2 < \ldots < y_n \), Algorithm 2 sets \( \hat{p}_i = i/n \), for \( 1 \leq i \leq n \). It follows that

\[
\hat{W}\hat{W}' = \begin{bmatrix} \sum_{i=1}^n \hat{p}_i^2 & \sum_{i=1}^n (1 - \hat{p}_i)\hat{p}_i \\ \sum_{i=1}^n (1 - \hat{p}_i)\hat{p}_i & \sum_{i=1}^n (1 - \hat{p}_i)^2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n \hat{p}_i^2 & \sum_{i=1}^n (1 - \hat{p}_i)\hat{p}_i \\ \sum_{i=1}^n (1 - \hat{p}_i)\hat{p}_i & \sum_{i=1}^n (1 - \hat{p}_i)^2 \end{bmatrix}.
\]

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 \( \hat{W}\hat{W}' \): First, \( \sum_{i=1}^n \hat{p}_i^2 = \frac{1}{n^2} \sum_{i=1}^n i^2 = \frac{n}{6}[1 + O(n^{-1})] \). Second, \( \sum_{i=1}^n (1 - \hat{p}_i)\hat{p}_i = \frac{1}{n^2} \sum_{i=1}^n i(n-i) = \frac{1}{n} \sum_{i=1}^n i - \frac{1}{n^2} \sum_{i=1}^n i^2 = \frac{n}{6}[1 + O(n^{-1})] \). Third, \( \sum_{i=1}^n (1 - \hat{p}_i)^2 = \frac{1}{n^2} \sum_{i=1}^n (n-i)^2 = \frac{1}{n^2} \sum_{i=0}^{n-1} i^2 = \frac{n}{9}[1 + O(n^{-1})] \). Combining them gives

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

Additionally, by direct calculations,

\[
n^{-1}(\hat{W}\hat{W}') = \begin{bmatrix} \frac{1}{n} \sum_i p_i \hat{p}_i & \frac{1}{n} \sum_i p_i (1 - \hat{p}_i) \\ \frac{1}{n} \sum_i (1 - p_i) \hat{p}_i & \frac{1}{n} \sum_i (1 - p_i)(1 - \hat{p}_i) \end{bmatrix} \tag{D.25}
\]
We now plug (D.24)-(D.25) into (D.23). It gives

\[
M = \begin{bmatrix}
1 & 1 \\
1 & -1 \\
\end{bmatrix}
\begin{bmatrix}
\frac{1}{n} \sum_i p_i \hat{p}_i & \frac{1}{n} \sum_i p_i (1 - \hat{p}_i) \\
\frac{1}{n} \sum_i (1 - p_i) \hat{p}_i & \frac{1}{n} \sum_i (1 - p_i) (1 - \hat{p}_i) \\
\end{bmatrix}
\begin{bmatrix}
4 & -2 \\
-2 & 4 \\
\end{bmatrix}
\begin{bmatrix}
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & -\frac{1}{2} \\
\end{bmatrix}
\]

\[
= \begin{bmatrix}
\frac{2}{n} \sum_i (p_i - \frac{1}{2}) & \frac{12}{n} \sum_i (p_i - \frac{1}{2}) (\hat{p}_i - \frac{1}{2}) \\
\end{bmatrix}
\]

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

\[
M = \begin{bmatrix}
1 & 0 \\
0 & \rho \\
\end{bmatrix} + O(n^{-1}).
\]

(D.26)

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

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

\[
|\hat{O}_{\pm,j} - O^*_\pm,j| \leq |\bar{O}_{\pm,j} - O^*_\pm,j| + |\bar{O}_{\pm,j}| \cdot \frac{1}{\|\bar{O}_\pm\|_1} - 1.
\]

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

\[
\hat{O}_{\pm,j} - O^*_\pm,j \leq |\bar{O}_{\pm,j} - O^*_\pm,j| + \frac{|\bar{O}_{\pm,j}|}{\|\bar{O}_\pm\|_1} \|\bar{O}_\pm - O^*_\pm\|_1.
\]

(D.27)

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

\[
\|\hat{O}_\pm - O^*_\pm\|_1 \leq 2\|\bar{O}_\pm - O^*_\pm\|_1.
\]

(D.28)

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

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

Write \( Z = \hat{D} - \hat{E}\hat{D} \). Then, \( \hat{D} = OW + Z \) and

\[
\hat{O} = (OW + Z)\hat{W}'(\hat{WW}')^{-1} = O^* + Z\hat{W}'(\hat{WW}')^{-1}.
\]

Let \( z_i \) be the \( i \)-th column of \( Z, 1 \leq i \leq n \). Plugging in the form of \( \hat{W} \), we have

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

37
It follows that

$$
\|\tilde{O}_{\pm,j} - O_{\pm,j}^*\|_1 \leq \max\left\{ \left\| \frac{1}{n} \sum_{i=1}^n \tilde{p}_i Z_{i,j} \right\|, \left\| \frac{1}{n} \sum_{i=1}^n (1 - \tilde{p}_i) Z_{i,j} \right\| \right\} \|n(WW')^{-1}\|_1
$$

$$
\leq C \max\left\{ \left\| \frac{1}{n} \sum_{i=1}^n \tilde{p}_i Z_{i,j} \right\|, \left\| \frac{1}{n} \sum_{i=1}^n (1 - \tilde{p}_i) Z_{i,j} \right\| \right\},
$$

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

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

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

$$
\left| \sum_{i=1}^n \tilde{p}_i Z_{i,j} \right| \leq C \sqrt{ns^{-1}F_j \log(m)} + Cs^{-1} \log(m) \leq C \sqrt{ns^{-1}F_j \log(m)},
$$

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

$$
\|\tilde{O}_{\pm,j} - O_{\pm,j}^*\|_1 \leq C \frac{\sqrt{F_j \log(m)}}{\sqrt{n}}.
$$

It follows from Cauchy-Schwarz inequality that

$$
\|\tilde{O} - O^*_+\|_1 \leq C \frac{\log(m)}{\sqrt{n}} \sum_{j \in S} \sqrt{F_j} \leq C \frac{\log(m)}{\sqrt{n}} \cdot |S|^{\frac{1}{2}} \left( \sum_{j \in S} F_j \right)^{\frac{1}{2}} \leq C \frac{|S| \log(m)}{\sqrt{n}}.
$$

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

\[ \square \]

**D.3 Proof of Theorem C.4**

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

We need some preparation. First, by our assumption, $F_j + \eta T_j = pO_{+j} + (1 - p)O_{-j} \geq c_1(O_{+j} + O_{-j}) = 2c_1 F_j$. Second, by (D.31) in the proof of Theorem C.3, $|\tilde{F}_j - F_j| \leq C \sqrt{F_j \log(m)/(ns)}$ and $|\tilde{T}_j - \rho T_j| \leq C \sqrt{F_j \log(m)/(ns)}$. Since $nsF_j \gg \log(m)$, we immediately obtain $|\tilde{F}_j - F_j| = O(F_j)$. Third, the condition (C.11) guarantees $n\theta^2 \rho^2 T_j^2 \geq \frac{\log^2(m)}{sF_j}$. In other words, $\rho|T_j| \gg \sqrt{F_j \log(m)/(ns)}$. 

38
So, \(|\hat{T}_j - \rho T_j| \ll \rho|T_j|\). We summarize these results as follows: for any \(j \in S\),

\[
\frac{|F_j + \eta T_j|}{F_j} \geq 2c_1, \quad \max\{\bar{F}_j - F_j, |\hat{T}_j - \rho T_j|\} \leq C \sqrt{\frac{\log(m)}{nsF_j}}, \quad \frac{|\hat{T}_j - \rho T_j|}{\rho|T_j|} = o(1). \tag{D.32}
\]

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

\[
|\hat{p} - p^*| = \frac{1}{2}|\hat{\eta} - \rho^{-1}\eta|. \tag{D.33}
\]

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

Suppose \(|\hat{\eta} - \rho^{-1}\eta| \leq c_1\). Let \(\hat{F} = \frac{1}{2}(\hat{O}_+ + \hat{O}_-)\) and \(\hat{T} = \frac{1}{2}('\hat{O}_+ - \hat{O}_-').\) Since \(p(1 - p) = (1 - \eta^2)/4\) and \(p\hat{O}_{+,j} + (1 - p)\hat{O}_{-,j} = \hat{F} + \eta\hat{T}_j\), the penalized MLE (A.3) has an equivalent form:

\[
\hat{\eta} = \arg\max_{\eta \in [-1,1]} \ell_{\lambda}(\eta), \quad \text{where } \ell_{\lambda}(\eta) \equiv s^{-1} \sum_{j \in S} d_j \log(\hat{F}_j + \eta \hat{T}_j) + \lambda \log(1 - \eta) + \lambda \log(1 + \eta).
\]

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

\[
s^{-1} \sum_{j \in S} d_j \log\left(1 + \frac{\hat{F}_j + \eta \hat{T}_j}{\hat{F}_j + \rho^{-1}\eta \hat{T}_j}\right) + \lambda \log\left(1 + \frac{\hat{\eta} - \rho^{-1}\eta}{1 + \rho^{-1}\eta}\right) + \lambda \log\left(1 - \frac{\hat{\eta} - \rho^{-1}\eta}{1 - \rho^{-1}\eta}\right) \geq 0. \tag{D.34}
\]

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

\[
\log\left(1 + \frac{\hat{\eta} - \rho^{-1}\eta}{1 + \rho^{-1}\eta}\right) + \log\left(1 - \frac{\hat{\eta} - \rho^{-1}\eta}{1 - \rho^{-1}\eta}\right) \\
\leq \frac{\hat{\eta} - \rho^{-1}\eta}{1 + \rho^{-1}\eta} - \frac{(\hat{\eta} - \rho^{-1}\eta)^2}{4(1 + \rho^{-1}\eta)^2} - \frac{\hat{\eta} - \rho^{-1}\eta}{1 - \rho^{-1}\eta} - \frac{(\hat{\eta} - \rho^{-1}\eta)^2}{4(1 - \rho^{-1}\eta)^2} \\
= - (\hat{\eta} - \rho^{-1}\eta) \frac{2\rho^{-1}\eta}{1 - \rho^{-2}\eta^2} - (\hat{\eta} - \rho^{-1}\eta)^2 \frac{1 + \rho^{-2}\eta^2}{2(1 - \rho^{-2}\eta^2)^2}. \tag{D.35}
\]

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

We thus have

\[
s^{-1} \sum_{j \in S} d_j \log\left(1 + \frac{\hat{\eta} - \rho^{-1}\eta}{\hat{F}_j + \rho^{-1}\eta \hat{T}_j}\right) \\
\leq (\hat{\eta} - \rho^{-1}\eta) \sum_{j \in S} \frac{s^{-1} d_j \hat{T}_j}{\hat{F}_j + \rho^{-1}\eta \hat{T}_j} - (\hat{\eta} - \rho^{-1}\eta)^2 \sum_{j \in S} \frac{s^{-1} d_j \hat{T}_j^2}{4(\hat{F}_j + \rho^{-1}\eta \hat{T}_j)^2}. \tag{D.36}
\]
We plug (D.35)-(D.36) into (D.34). It gives

\[(\hat{\eta} - \rho^{-1}\eta)X_1 - (\hat{\eta} - \rho^{-1}\eta)^2 X_2 \geq 0, \quad \implies |\hat{\eta} - \rho^{-1}\eta| \leq \frac{|X_1|}{X_2}, \quad (D.37)\]

where

\[X_1 = \sum_{j \in S} \frac{s^{-1}d_j \hat{T}_j}{F_j + \rho^{-1}\eta \hat{T}_j} - \frac{2\lambda \rho^{-1}\eta}{1 - \rho^{-2}\eta^2}, \quad X_2 = \sum_{j \in S} \frac{s^{-1}d_j \hat{T}_j^2}{4(F_j + \rho^{-1}\eta \hat{T}_j)^2} + \frac{\lambda(1 + \rho^{-2}\eta^2)}{2(1 - \rho^{-2}\eta^2)^2}.\]

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

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

\[
\left| \sum_{j \in S} \frac{\hat{T}_j s^{-1}E_{d_j}}{F_j + \rho^{-1}\eta \hat{T}_j} \right| \\
\leq \sum_{j \in S} \frac{\rho T_j s^{-1}E_{d_j}}{F_j + \eta \hat{T}_j} + \left| \sum_{j \in S} \frac{(\hat{T}_j - \rho T_j) s^{-1}E_{d_j}}{F_j + \rho^{-1}\eta \hat{T}_j} \right| + \left| \sum_{j \in S} \rho T_j s^{-1}E_{d_j} \left( \frac{1}{F_j + \rho^{-1}\eta \hat{T}_j} - \frac{1}{F_j + \eta \hat{T}_j} \right) \right| \\
\leq \sum_{j \in S} \frac{\rho T_j s^{-1}E_{d_j}}{F_j + \eta \hat{T}_j} + \left| \sum_{j \in S} \frac{\hat{T}_j - \rho T_j |s^{-1}E_{d_j}|}{2(F_j + \eta \hat{T}_j)} \right| + \sum_{j \in S} \rho T_j |s^{-1}E_{d_j}| \frac{|\hat{T}_j - F_j| + \rho^{-1}\eta |\hat{T}_j - T_j|}{2(F_j + \eta \hat{T}_j)} \\
\leq \left| \sum_{j \in S} T_j \right| + \frac{1}{2} \sum_{j \in S} |\hat{T}_j - T_j| + \frac{1}{2} \sum_{j \in S} \rho |T_j| (|\hat{T}_j - F_j| + \rho^{-1}\eta |\hat{T}_j - T_j|) \\
\leq 0 + C\|\hat{F} - F\|_1 + C\|\hat{T} - T\|_1 \\
\leq C \sqrt{\frac{|S| \log(m)}{ns}},
\]

where the second last line is due to $\sum_{j \in S} O_{+,j} = \sum_{j \in S} O_{-,j} = 1$ and the last line is by Theorem C.3.

Moreover, since the covariance matrix of $d_j$ is $s \cdot \text{diag}(F + \eta T) - s(F + \eta T)(F + \eta T)' \preceq s \cdot \text{diag}(F + \eta T)$, we have

\[
\text{Var} \left( \sum_{j \in S} \frac{\hat{T}_j s^{-1}d_j}{F_j + \rho^{-1}\eta \hat{T}_j} \right) \leq \sum_{j \in S} \frac{\hat{T}_j^2 s^{-2} \cdot s(F_j + \eta T_j)}{(F_j + \rho^{-1}\eta \hat{T}_j)^2} \\
\leq C s^{-1} \sum_{j \in S} \frac{\rho^2 T_j^2}{F_j} + C s^{-1} \sum_{j \in S} \frac{\hat{T}_j - \rho T_j)^2}{F_j} \\
\leq C s^{-1} \rho^2 \Theta + C s^{-1} \frac{|S| \log(m)}{ns},
\]

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

\[
\sum_{j \in S} \frac{\hat{T}_j s^{-1}d_j}{F_j + \rho^{-1}\eta \hat{T}_j} \overset{(d)}{=} \sum_{\ell=1}^s \xi_\ell, \quad \text{with} \quad \xi_\ell \equiv \sum_{j \in S} \frac{b_{\ell,j} \hat{T}_j}{F_j + \rho^{-1}\eta \hat{T}_j}.
\]
Conditioning on \((\hat{F}, \hat{T})\), \(\{\xi_\ell\}_{\ell=1}^{s}\) 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 \(|\sum_{\ell=1}^{s}\xi_\ell|\) and \(\text{Var}(\sum_{\ell=1}^{s}\xi_\ell)\). We apply the Bernstein’s inequality and find out that, for any \(\epsilon \in (0, 1)\), with probability \(1 - \epsilon\),

\[
\left| \sum_{j \in S} \frac{\hat{T}_{j}s^{-1}d_j}{\hat{F}_j + \rho^{-1}\eta \hat{T}_j} \right| \leq C\sqrt{\frac{|S|\log(m)}{ns}} + C\rho\sqrt{\frac{\Theta \log(\epsilon^{-1})}{s}} + \rho \log(\epsilon^{-1})\frac{2c_1s}{s} 
\]

where the last line is because \(s\Theta \to \infty\). We plug (D.38) into the expression of \(X\). Additionally, we notice that
\[
1 - \rho^{-2}\eta^{-2} = (1 + \rho^{-1}\eta)(1 - \rho^{-1}\eta) = 4p^*(1 - p^*) \geq 4c_2. 
\]

Hence, with probability \(1 - \epsilon\),

\[
|X| \leq \frac{\lambda}{2c_1} |\rho^{-1}\eta| + C\sqrt{\frac{|S|\log(m)}{ns}} + C\rho\sqrt{\frac{\Theta \log(\epsilon^{-1})}{s}}. 
\]

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

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

At the same time,

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

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

\[
\sum_{j \in S} \frac{\hat{T}_j^2s^{-1}d_j}{4(\hat{F}_j + \rho^{-1}\eta \hat{T}_j)^2} \geq C^{-1} \rho^2 \Theta. 
\]

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

\[
X_2 \geq \frac{\lambda}{2} + C^{-1} \rho^2 \Theta. 
\]

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

\[
|\hat{\eta} - \rho^{-1}\eta| \leq C \frac{\lambda|\rho^{-1}\eta| + \sqrt{\frac{|S|\log(m)}{ns}} + \rho \sqrt{\frac{\Theta \log(\epsilon^{-1})}{s}}}{\lambda + \rho^2 \Theta}
\]
By separating two cases, \( \lambda \leq \rho^2 \Theta \) and \( \lambda > \rho^2 \Theta \), we immediately obtain

\[
|\hat{\eta} - \rho^{-1} \eta| \leq C \begin{cases} \frac{\lambda}{\rho^2 \Theta} |\rho^{-1} \eta| + \left(\frac{\sqrt{|S| \log(m)}}{\rho^2 \Theta \sqrt{n s}} + \frac{\sqrt{\log(e^{-1})}}{\rho^2 \Theta s}\right), & \text{if } \lambda \leq \rho^2 \Theta, \\ |\rho^{-1} \eta| + \frac{\rho^2 \Theta}{\lambda} \left(\frac{\sqrt{|S| \log(m)}}{\rho^2 \Theta \sqrt{n s}} + \frac{\sqrt{\log(e^{-1})}}{\rho^2 \Theta s}\right), & \text{if } \lambda > \rho^2 \Theta. \end{cases}
\]

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

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

\[
\ell_\lambda'(\eta) = \sum_{j \in S} \frac{d_j T_j}{F_j + \eta T_j} - \frac{2\lambda \eta}{1 - \eta^2}, \quad \ell_\lambda''(\eta) = -\sum_{j \in S} \frac{d_j T_j^2}{2(F_j + \eta T_j)^2} - \frac{\lambda(1 + \eta^2)}{2(1 - \eta^2)^2}.
\]

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

\[
\ell_\lambda'(1 - 2c_1) = \sum_{j \in S} \frac{\rho T_j(F_j + \eta T_j)}{F_j + (1 - 1.9c_1)\rho T_j} - \frac{2\lambda(1 - 1.9c_1)}{[1 - (1 - 1.9c_1)^2]^2} + O\left(\sqrt{\frac{|S| \log(m)}{n s}}\right) + O\left(\rho \sqrt{\frac{\max(\log(e^{-1}))}{s}}\right)
\]

\[
= -\sum_{j \in S} \frac{\rho^2([1 - 1.9c_1] - \rho^{-1} \eta T_j^2)}{[F_j + (1 - 1.9c_1)\rho T_j]^2} - \frac{2\lambda(1 - 1.9c_1)}{[1 - (1 - 1.9c_1)^2]^2} + o(\lambda + \rho^2 \Theta)
\]

\[
\geq -0.1c_1 \rho^2 \Theta - \frac{2\lambda(1 - 1.9c_1)}{[1 - (1 - 1.9c_1)^2]^2} + O\left(\sqrt{\frac{|S| \log(m)}{n s}}\right) + o(\lambda + \rho^2 \Theta).
\]

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 \(\pm 1\) by at least 1.9c_1. For any \( \eta_0 \) in this interval, \( F_j + \eta_0 T_j \geq cF_j \) for a constant \( c > 0 \). This allows us to mimic the proof of (D.40)-(D.41) to get

\[-\ell_\lambda''(\eta_0) \geq C^{-1}(\lambda + \rho^2 \Theta), \quad \text{for } \eta_0 \text{ in this compact interval.}\]

By Taylor expansion, there exists \( \eta_0 \), whose value is between \( \rho^{-1} \eta \) and \( \hat{\eta} \), such that

\[0 = \ell_\lambda'(\hat{\eta}) = \ell_\lambda'(\rho^{-1} \eta) + \ell_\lambda''(\eta_0)(\hat{\eta} - \rho^{-1} \eta).\]

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