Beyond Factors: AI on SEC Text

Daniel Buckman, Santiago Etchepare, Michael Logan, Livia Mucciolo
The University of Chicago Booth School of Business

1 Introduction

Public companies in the United States are required to promptly disclose major corporate events through Form 8-K filings, which announce unscheduled material events or corporate changes outside the periodic reporting cycle. These 8-K "current reports" serve as a key channel for timely information release and have an immediate impact on stock prices Zhao [2017]. Unlike annual 10-K or quarterly 10-Q reports that encompass a broad range of information (much of it anticipated or repetitive), 8-K filings focus on specific material events – such as mergers and acquisitions, earnings surprises, leadership changes, or other significant developments – thereby offering a higher signal-to-noise ratio in their textual content. In other words, because each 8-K is triggered by a noteworthy event, the information density per filing tends to be higher relative to routine filings.

Recognizing this high information content, our study focuses on 8-K filings as an attractive source of textual signals for predicting stock returns. Recent advances in artificial intelligence (AI) – particularly in Natural Language Processing (NLP) – enable more sophisticated analysis of financial text than was previously possible. Traditional approaches to textual analysis in finance often relied on simple metrics (e.g. word counts or sentiment dictionaries) or basic machine learning classifiers (e.g. naive Bayes, logistic regressions) to quantify the tone or content of disclosures.

While such approaches have shown that textual information from sources like SEC filings and news media can predict aspects of firm performance and stock returns, they have limitations in capturing context and complex linguistic patterns. In contrast, modern transformer-based language models and related deep learning architectures can learn rich semantic representations from unstructured text, potentially yielding more powerful predictive signals. We therefore propose to train a transformer-based AI model on the textual content of 8-K filings, with the objective of forecasting subsequent stock return movements. By leveraging a state-of-the-art NLP model, we aim to determine whether nuanced language cues in 8-K disclosures contain incremental information for stock return prediction beyond what traditional financial factors capture.

A central question we investigate is whether the predictive signals extracted from 8-K texts represent genuine alpha – that is, excess return not explained by standard risk premia. To test this, we evaluate the model's performance in the context of well-known asset pricing benchmarks. Specifically, after generating return forecasts or trading signals from the 8-K-based transformer model, we assess their performance controlling for Fama–French risk factors (Fama and French [1993], Fama and French [2015]). This involves examining risk-adjusted returns (e.g. intercepts from regressions on the Fama–French five factors) to see if any outperformance remains. By doing so, we directly address whether the information in textual disclosures can yield abnormal returns or whether it merely proxies for known risk exposures.

The structure of this paper is as follows. Section 2 reviews the relevant literature on financial text analysis, highlighting the progression from traditional sentiment metrics to modern transformer-based models and their application in capital markets research. Section 3 describes the dataset, including the sourcing and preprocessing of 8-K filings and the construction of firm-level return metrics. Section 4 outlines our methodology, detailing the architecture of the transformer model, the training procedure, and the framework used to generate and evaluate return forecasts. Section 5 presents the empirical results, focusing on the statistical significance of our intercepts and a range of robustness checks. Section 6 concludes by discussing the implications of our findings for both asset pricing theory and the practical use of AI-driven text analysis.

2 Literature Review

Recent advancements in ML have invigorated research into the predictive power of financial disclosures, particularly textual information embedded in corporate filings and financial news. The intersection of NLP and asset pricing has opened new avenues for understanding how qualitative data contributes to market outcomes. Early studies, such as Tetlock [2007], utilized sentiment dictionaries to show that pessimism in Wall Street Journal columns had modest predictive power for short-term stock price declines. This line of work was extended by Loughran and McDonald [2011], who introduced a domain-specific sentiment dictionary for financial texts and demonstrated that negative language in 10-K filings is associated with immediate market reactions. Their findings emphasized the importance of tailoring sentiment tools to financial language rather than relying on general-purpose lexicons.

Textual information was further analyzed for its forward-looking implications. Li [2010] used a Naïve Bayes classifier to analyze the Management Discussion and Analysis (MD&A) section of filings, showing that optimistic forward-looking statements are positively associated with future earnings. His work illustrated that when linguistic tone aligns with underlying fundamentals, it can help explain variations in future firm performance. Topic modeling has also been deployed to extract deeper insights from disclosures. Kogan et al. [2009] applied Latent Dirichlet Allocation (LDA) to SEC fillings and found that certain topics, such as those relating to technology and competition, predicted future return volatility.

As computational capabilities grew, the financial literature increasingly embraced ML approaches to handle the high dimensionality and complexity of text. Antweiler and Frank [2004] were among the first to apply machine learning to textual sentiment, using a Naïve Bayes model to study stock message boards. However, early models struggled with context sensitivity. The introduction of transformer architectures, particularly BERT by Devlin et al. [2019], marked a major turning point in NLP. These models learn context-aware embeddings that significantly enhance text understanding. Financial applications soon followed, with domain-specific variants like FinBERT [Huang et al., 2022] trained on financial corpora such as 10-K, 10-Q filings, and 8-K. FinBERT has outperformed traditional sentiment methods in classifying financial text, including analyst reports, due to its ability to interpret complex financial phrases.

Recent studies have combined these deep learning models with asset pricing frameworks. For instance, Lopez-Lira [2019] used ML to extract interpretable risk factors from 10-K risk disclosures and demonstrated that these text-derived risks explain cross-sectional returns better than traditional factor models. By constructing portfolios based on firms' exposure to these risks—technology, demand, production, and international exposure—he proposed a new four-factor model that complements or even outperforms the Fama-French benchmarks.

Broader applications of ML in finance are illustrated by Gu et al. [2020], who evaluated a wide range of ML algorithms—including deep neural networks—on traditional numeric predictors. They found that non-linear models can uncover predictive structures in the data that linear models overlook, with significantly higher out-of-sample R^2. The ability of ML to extract value from both structured and unstructured data has led to a surge in studies combining numeric and text features. Recurrent neural networks (RNNs), long short-term memory (LSTM) models, and transformer-based architectures have been deployed on earnings call transcripts and MD&A sections to capture sequential dependencies and tone changes, particularly in forward-looking content. Although early RNN-based models had limited success, transformer models have proven more adept at focusing on the most informative parts of long documents.

The progression from word counts to deep contextual models underscores a broader trend in financial research: the shift from simple statistical correlations to more sophisticated, nonlinear predictive modeling. As pretrained models like FinBERT continue to evolve and as computing costs decrease, the use of large language models in asset pricing and financial forecasting is becoming increasingly prevalent.

3 Data

3.1 SEC Filings and Textual Data

Our study centers on Form 8-K filings submitted by U.S. public companies, specifically those included in the 2025 S&P 500 index. Form 8-K, mandated by the U.S. Securities and Exchange Commission (SEC), is a current report that companies must file to disclose unscheduled material events or corporate changes deemed important to shareholders. Unlike Form 10-K or Form 10-Q, which follow a regular reporting cycle and cover a broad set of financial metrics and disclosures, the 8-K is event-driven and typically focuses on a single, significant development.

These filings cover a wide range of material events, such as earnings announcements, mergers and acquisitions, executive transitions, changes in control, auditor replacements, credit agreements, and legal proceedings. Because each filing is triggered by a noteworthy occurrence, the text is often more information-dense and time-sensitive than routine regulatory filings. The market frequently reacts to such disclosures in real time, making them a valuable source of high-frequency textual signals.

We restrict our analysis to firms that are constituents of the S&P 500 as of January 2025, ensuring a consistent and liquid cross-section of mature public firms. The sample period spans from January 1, 2000 to January 1, 2025, capturing 25 years of 8-K filings for these firms. This extended period allows us to evaluate model performance across multiple business cycles, regulatory changes, and technology regimes.

All 8-K filings are sourced from the SEC's EDGAR database, and we extract the primary narrative content from each filing, excluding exhibits and non-substantive boilerplate. Filings are preprocessed to remove HTML tags, legal disclaimers, and formatting noise.

3.2 Stock Returns and Factor Data

Closing prices and market capitalizations for all firms are obtained from Wharton Research Data Services (WRDS), which offers comprehensive and reliable historical coverage. These prices are used to calculate realized returns over various post-filing windows and to compute market-level returns by aggregating across market capitalizations.

To assess risk-adjusted performance, we incorporate the Fama–French five-factor model as a benchmark. This model includes: (1) the market excess return (MKT), (2) the size premium (SMB: small minus big), (3) the value premium (HML: high book-to-market minus low), (4) the profitability factor (RMW: robust minus weak), and (5) the investment factor (CMA: conservative minus aggressive). Daily and monthly values for these factors are also downloaded from WRDS Fama–French database, covering the full sample period. These factors allow us to isolate whether the return forecasts generated by our model reflect genuine alpha or are merely exposures to known sources of systematic risk.

4 Methodology

4.1 Text Preprocessing and Vectorization

After obtaining and cleaning the filing texts as described, the next step is to convert each document's text into a numerical representation that a machine learning model can ingest—a process often called vectorization. Given the advances in NLP, we prioritize using pretrained language model embeddings for this task. In practice, this involves leveraging models like FinBERT or BERT to encode each filing's text into a fixed-length vector (or a set of vectors). These models are pretrained on large corpora and can capture nuanced financial language, tone, and context.

One straightforward approach is to use the pretrained model to generate a document-level embedding. For example, we can feed the text through FinBERT, a BERT-based model fine-tuned on financial documents like 10-K and 10-Q filings. FinBERT produces contextual embeddings for each token and a special [CLS] token that can act as an aggregate representation of the entire input. If the filing text exceeds the token limit (typically 512 tokens), we split it into manageable chunks (such as paragraphs or sections), embed each chunk separately, and then aggregate the chunk-level embeddings—either by averaging or using more sophisticated methods like attention pooling—into a single vector.

We also include Nomic embeddings as an alternative approach, given their strong performance across a wide range of language tasks. By comparing outputs across different embedding models, we aim to test the robustness of our results. The goal is to ensure that the performance of downstream tasks reflects real semantic signal capture, rather than artifacts from any specific model.

4.2 Modeling Approach

Let $\{x_i, r_i\}_{i=1}^{N}$ denote a dataset of N samples, where each $x_i \in \mathbb{R}^{T\times d}$ is a sequence of T dense vector embeddings derived from a U.S. SEC Form 8-K filing, and $r_i \in \mathbb{R}$ is the realized return at horizon $h \in \{1, 30\}$ days. The objective is to learn a predictive function $f_\theta : \mathbb{R}^{T\times d} \to \mathbb{R}$, parameterized by θ, that minimizes the empirical mean squared error.

Model Architecture

The predictive function f_θ is implemented as a transformer-based model composed of the following components:

(a) **Input Projection:** A linear projection maps the input embedding to the transformer's hidden dimension:

$$z_i^{(0)} = W_{\text{proj}} x_i + b_{\text{proj}}, \quad W_{\text{proj}} \in \mathbb{R}^{h\times d},\ b_{\text{proj}} \in \mathbb{R}^{h}$$

(b) **Transformer Encoder:** The projected vector is treated as a sequence of length one and passed through L stacked Transformer encoder layers. For each layer $\ell = 1, \ldots, L$:

$$\begin{aligned} \text{head}_j &= \text{Attention}(zW_j^Q, zW_j^K, zW_j^V), \quad j = 1, \ldots, H \\ \text{MultiHead}(z) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_H)W^O \\ z_i^{(\ell)} &= \text{LayerNorm}\left(z_i^{(\ell-1)} + \text{MultiHead}(z_i^{(\ell-1)})\right) \\ z_i^{(\ell)} &= \text{LayerNorm}\left(z_i^{(\ell)} + \text{FFN}(z_i^{(\ell)})\right) \end{aligned}$$

where $\text{FFN}(z) = \phi(zW_1 + b_1)W_2 + b_2$ and ϕ is the GELU activation.

(c) **Global Pooling:** A mean pooling operation reduces the sequence output to a fixed-length representation:

$$\tilde{z}_i = \frac{1}{T}\sum_{t=1}^{T} z_i^{(L)}[t] \in \mathbb{R}^h$$

(d) **Residual Feedforward Block:** The pooled representation is passed through a residual feedforward block:

$$h_i = \text{LayerNorm}\left(\tilde{z}_i + \phi(\tilde{z}_i W_{\text{ff1}} + b_{\text{ff1}})W_{\text{ff2}} + b_{\text{ff2}}\right)$$

(e) **Output Layer:** A final linear transformation outputs the predicted return:

$$\hat{r}_i = h_i^\top w_{\text{out}} + b_{\text{out}}, \quad w_{\text{out}} \in \mathbb{R}^h, \; b_{\text{out}} \in \mathbb{R}$$

The full model prediction is given by:

$$f_\theta(x_i) := \hat{r}_i = \left[\text{LayerNorm}\left(\tilde{z}_i + \phi(\tilde{z}_i W_{\text{ff1}} + b_{\text{ff1}})W_{\text{ff2}} + b_{\text{ff2}}\right)\right]^\top w_{\text{out}} + b_{\text{out}}$$

Optimization Procedure

The model is trained by minimizing the loss:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(r_i - f_\theta(x_i))^2$$

using the AdamW optimizer with learning rate $\eta > 0$ and L2 weight decay $\lambda > 0$. Each parameter θ_t is updated at iteration t via:

$$\theta_{t+1} = \theta_t - \eta\left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda\theta_t\right)$$

where $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected estimates of first and second moments. Gradient clipping is applied with a maximum norm constraint, and training includes early stopping based on validation loss.

4.3 Measuring Excess Returns

Predictive Return Modeling

We train a model to forecast returns, daily and monthly. Let X_t be features at time t (Vectorized 8-K forms), and r_{t+1} the return.

$$\hat{r}_{t+1} = f(X_t; \theta)$$

Model training minimizes:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \left(r_{t+1}^{(i)} - \hat{r}_{t+1}^{(i)} \right)^2$$

Fama-French 5-Factor Regression

The predicted market weight aggregated returns $\hat{r}_{t+1}^{agg}$ are regressed on the Fama-French 5 factors:

$$\hat{r}_{t+1}^{agg} = \alpha + \beta_1 MKT_{t+1} + \beta_2 SMB_{t+1} + \beta_3 HML_{t+1} + \beta_4 RMW_{t+1} + \beta_5 CMA_{t+1} + \varepsilon_t$$

Evaluating the Intercept (α)

The intercept reflects abnormal returns unexplained by the factors. We test:

$$H_0 : \alpha = 0 \quad \text{vs.} \quad H_1 : \alpha \neq 0$$

Significant α suggests the model captures returns beyond known systematic risk.

5 Results

5.1 Main Findings

Figure 1 displays the training and validation loss curves for both 1-day and 30-day return prediction tasks. In both cases, the model converges quickly and shows minimal overfitting, with validation loss tracking training loss closely. This indicates that the transformer architecture generalizes well on unseen samples.

To assess whether the model's predictions contain excess return (alpha) beyond known risk factors, we regress predicted returns on the Fama-French five-factor model plus momentum and the risk-free rate. The OLS regression on 1-day predicted returns (Figure 2)

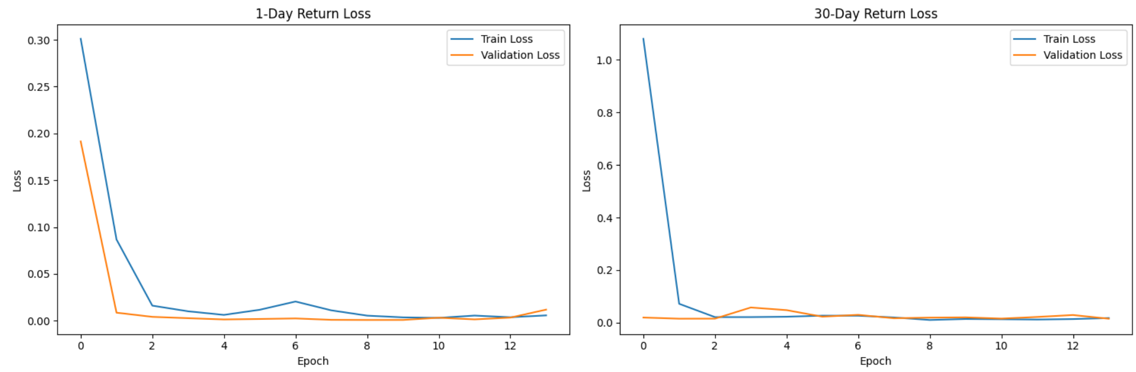

Figure 1: Training and validation loss over epochs

shows a statistically significant intercept of 0.0002 ($p < 0.001$), which annualizes to approximately 5%—a strong indication of short-term alpha. Notably, none of the factor loadings are statistically significant, suggesting that the model's signal is not merely proxying for existing systematic risk factors. It delivers a lower R^2 (1.2%) than the 30-day regression, but that's expected: daily returns are dominated by noise, whereas established risk factors have time to assert their influence over a month. What matters is that the 1-day regression still reports a statistically significant alpha, confirming our model is capturing signals beyond the standard factors.

			OLS Regression Result			
Dep. Variable:		weighted_1d_pred	R-squared:			0.012
Model:		OLS	Adj. R-squared:			0.01
Method:		Least Squares	F-statistic:			6.385
No. Observations:		3634	Prob (F-statistic):			1.75E-07
Df Residuals:		3626	Log-Likelihood:			24075
Df Model:		7	AIC:			-4.81E+04
			BIC:			-4.81E+04
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.0002	5.00E-05	4.011	0.000	0.000	0.000
mktrf	0.0002	0.001	0.397	0.691	-0.001	0.001
smb	0.0003	0.001	0.303	0.762	-0.002	0.002
hml	-0.0005	0.001	-0.537	0.591	-0.002	0.001
rmw	-0.0020	0.001	-1.589	0.112	-0.004	0.000
cma	0.0008	0.002	0.490	0.624	-0.002	0.004
umd	-0.0010	0.001	-1.715	0.086	-0.002	0.000
rf	-0.4570	0.076	-6.006	0.000	-0.606	(0.308)
Omnibus:		310.798	Durbin-Watson:			1.739
Prob(Omnibus):		0.000	Jarque-Bera (JB):			1755.631
Skew:		-0.176	Prob(JB):			0.000
Kurtosis:		6.387	Cond. No.			1.43E+04

Figure 2: OLS regression of 1-day predicted returns on Fama-French factors

In contrast, the regression on 30-day predicted returns (Figure 3) yields an insignificant intercept ($p = 0.708$), with a low R^2 of 6.4%. This suggests that while our model extracts meaningful predictive signals from 8-K disclosures in the short term, this alpha decays over longer horizons. The lack of significant factor exposure remains consistent, reinforcing the independence of the textual signal from standard asset pricing factors.

OLS Regression Result						
Dep. Variable:		weighted_30d_pred	R-squared:			0.064
Model:		OLS	Adj. R-squared:			0.034
Method:		Least Squares	F-statistic:			2.132
No. Observations:		225	Prob (F-statistic):			0.0415
Df Residuals:		217	Log-Likelihood:			888.03
Df Model:		7	AIC:			-1760
			BIC:			-1733
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0090	0.024	0.375	0.708	-0.038	0.056
mktrf	0.0156	0.008	1.950	0.052	0.000	0.031
smb	0.0102	0.014	0.717	0.474	-0.018	0.038
hml	0.0170	0.014	1.204	0.230	-0.011	0.045
rmw	-0.0008	0.018	-0.044	0.965	-0.037	0.035
cma	-0.0023	0.022	-0.106	0.915	-0.045	0.041
umd	0.0129	0.008	1.572	0.117	-0.003	0.029
rf	-0.4502	0.207	-2.173	0.031	-0.859	-0.042
Omnibus:		22.268	Durbin-Watson:			1.689
Prob(Omnibus):		0.000	Jarque-Bera (JB):			58.231
Skew:		-0.389	Prob(JB):			2.27E-13
Kurtosis:		5.367	Cond. No.			653

Figure 3: OLS regression of 30-day predicted returns on Fama-French factors

5.2 Robustness Check

The model's performance is robust across different training configurations. The optimization history in Figure 4 demonstrates convergence to a consistent objective value across trials, indicating stable hyperparameter tuning.

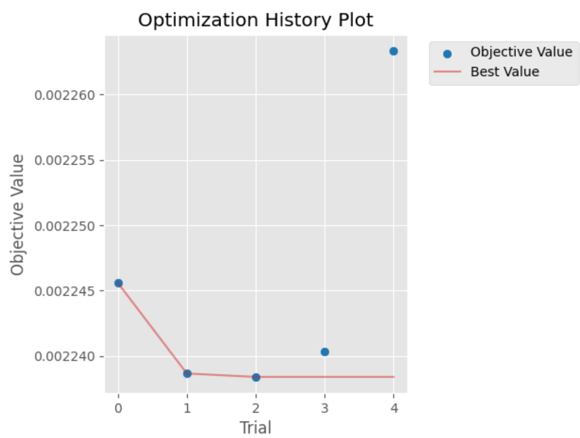

Figure 4: Hyperparameter optimization history

Moreover, five-fold cross-validation confirms that the 1-day return predictions generalize well, with average validation loss remaining low across folds (Figure 5). This consistency across splits further supports the presence of a short-term alpha signal. For the 30-day horizon, the model's performance is less stable and less predictive, consistent with the absence of statistically significant alpha in the corresponding OLS regression.

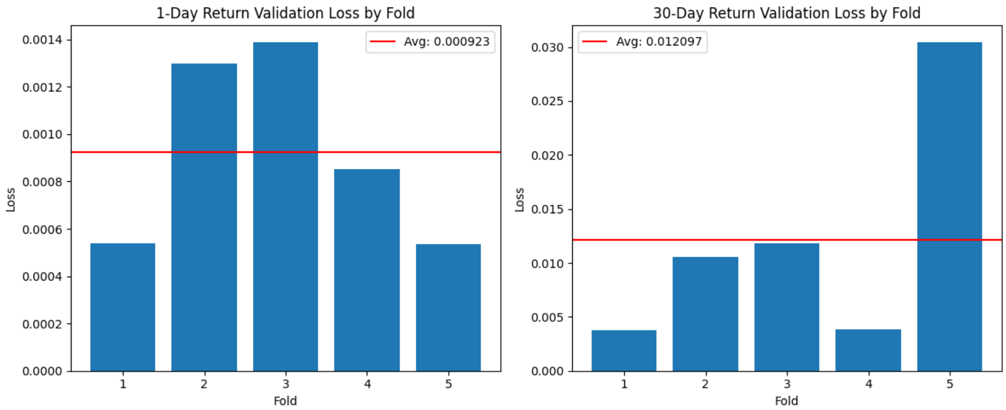

Figure 5: Cross-validation loss across folds for 1-day and 30-day horizons

6 Conclusions

This paper demonstrates that transformer-based models trained on 8-K filings can extract meaningful short-term predictive signals for stock returns. Specifically, we identify a statistically significant alpha in 1-day return forecasts, which annualizes to approximately 5%, and appears independent of standard Fama-French risk factors. However, this signal does not persist at the 30-day horizon, where alpha becomes statistically insignificant and explanatory power diminishes.

These findings suggest that the textual content of 8-K disclosures contains short-lived information that markets may not immediately and fully incorporate. While the signal decays rapidly, it presents a window of opportunity for short-term strategies. Our results underscore the value of modern NLP techniques in asset pricing and open avenues for further exploration into high-frequency financial text analysis. While we did find significant signal, it is unclear to what extent this can be implemented due to trading costs inherent in such short-term strategies.

References

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59(3), 1259–1294.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1–22.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3), 535–574.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223–2273.

Huang, A. H., Wang, H., & Yang, Y. (2022). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 39(4), 2478–2505.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009). Predicting risk from financial reports with regression. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 272–280.

Li, F. (2010). The information content of forward-looking statements in corporate filings: A naïve Bayesian machine learning approach. *Journal of Accounting Research*, 48(5), 1049–1102.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35–65.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187–1230.

Lopez-Lira, A. (2019). Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns. *Working Paper*, University of Pennsylvania. Retrieved from Mack Institute for Innovation Management website.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139–1168.

Zhao, X. (2017). Does information intensity matter for stock returns? Evidence from Form 8-K filings. *Management Science*, 63(5), 1382–1404.