

Risky Words and Returns

Sina Seyfi



Risks and Returns

Risks and Returns

$$\mathbb{E}(r) = f(\text{risk})$$


Risks and Returns

$$\mathbb{E}(r) = f(\text{risk})$$

Risk

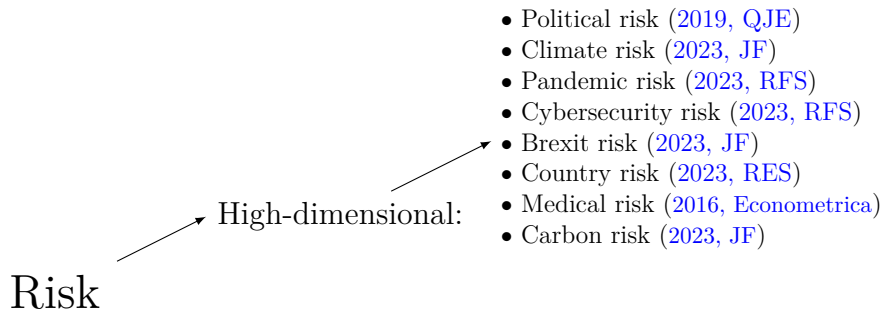
Risks and Returns

$$\mathbb{E}(r) = f(\text{risk})$$

Risk  High-dimensional:

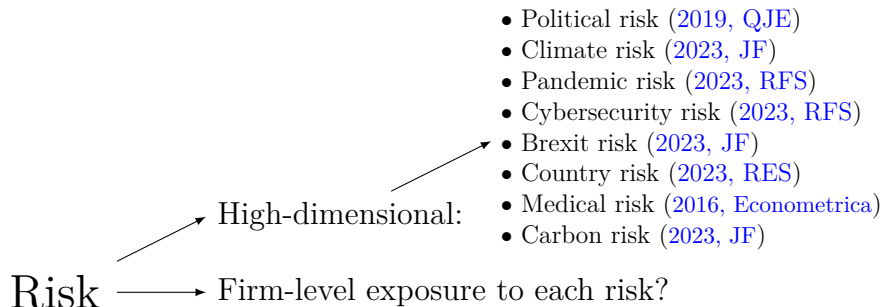
Risks and Returns

$$\mathbb{E}(r) = f(\text{risk})$$



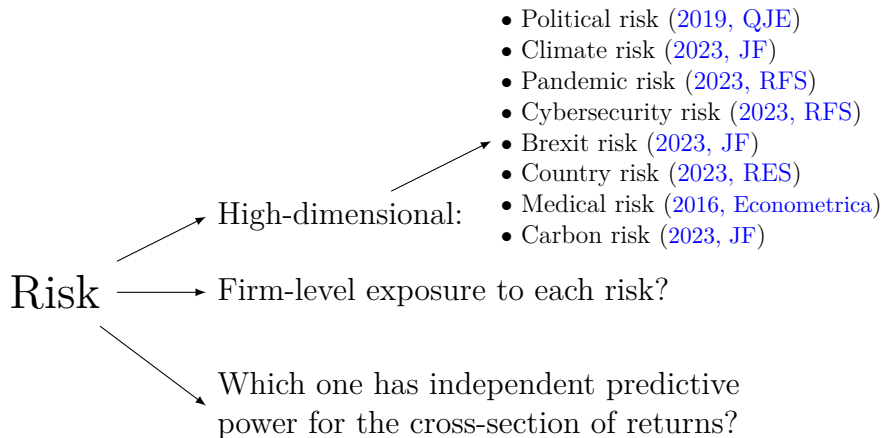
Risks and Returns

$$\mathbb{E}(r) = f(\text{risk})$$



Risks and Returns

$$\mathbb{E}(r) = f(\text{risk})$$



$$\mathbb{E}(r_{i,t}) = f(\text{risk disclosures}_{i,t-1})$$

$$\mathbb{E}(r_{i,t}) = f(\text{risk disclosures}_{i,t-1})$$

I show that:

- ▶ Strong return predictability
 - ▶ Up to 20% annual alpha 2005-2023

$$\mathbb{E}(r_{i,t}) = f(\text{risk disclosures}_{i,t-1})$$

I show that:

- ▶ Strong return predictability
 - ▶ Up to 20% annual alpha 2005-2023
- ▶ Sources of risks \rightarrow jointly predict returns

$$\mathbb{E}(r_{i,t}) = f(\text{risk disclosures}_{i,t-1})$$

I show that:

- ▶ Strong return predictability
 - ▶ Up to 20% annual alpha 2005-2023
- ▶ Sources of risks → jointly predict returns
 - ▶ Energy
 - ▶ Loan
 - ▶ Clients
 - ▶ Manufacturing
 - ▶ Company
 - ▶ Income
 - ▶ Drug



1.75 %



1.75 %



-1.50%

Risk sections

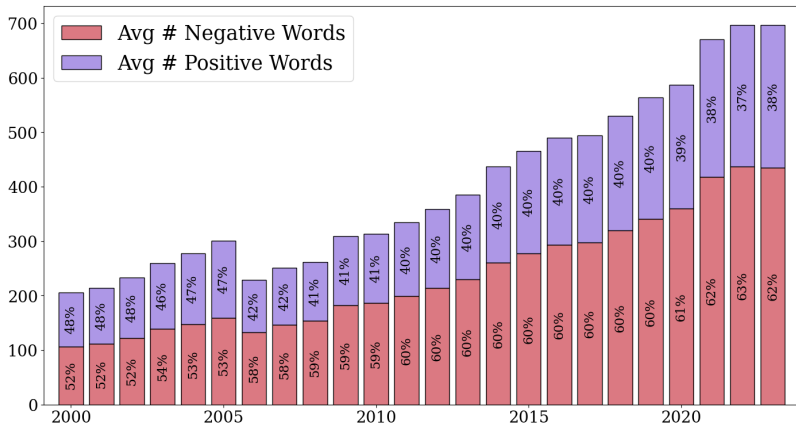


Figure: The average number of positive and negative words in risk sections per year



→ 0.5%



text_i : vector of text

0.5%



text_i : vector of text

0.5%

$$\mathbb{E}(r_{i,t}) = f(\text{text}_{i,t-1})$$



text_i : vector of text

0.5%

$$\mathbb{E}(r_{i,t}) = f(\text{text}_{i,t-1})$$

$$r_{i,t} = \alpha_t + \beta_t \text{text}_{i,t-1} + \epsilon_{i,t}$$

TF-IDF (Term Frequency-Inverse Document Frequency)

Firm 1: Covid hurts.

Firm 2: Carbon emission and technology matter.

Firm 3: Climate change and Covid affect us.

	Covid	hurts	carbon	emission	technology	matter	climate	change	affect	us	and
Firm 1	0.34	0.94	0	0	0	0	0	0	0	0	0
Firm 2	0	0	0.49	0.49	0.49	0.49	0	0	0	0	0.18
Firm 3	0.19	0	0	0	0	0	0.48	0.48	0.48	0.48	0.19

TF-IDF



TF-IDF =

$$\begin{matrix} & \text{word1} & \text{word2} & \text{.....} & \text{word N} \\ \left(\begin{array}{cccc} & & \dots & \\ & & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & \end{array} \right) \end{matrix}$$

TF-IDF



$$\text{TF-IDF} = \begin{matrix} & \text{word1} & \text{word2} & \text{.....} & \text{word N} \\ \begin{pmatrix} & & & & \\ & & \dots & & \\ & & \dots & & \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ & & \dots & & \end{pmatrix} \end{matrix}$$

- ▶ Predict expected returns using TF-IDF
 - ▶ words \rightarrow expected returns

TF-IDF



$$\Rightarrow \text{TF-IDF} = \begin{pmatrix} & \text{word1} & \text{word2} & \dots & \text{word N} \\ & & & \dots & \\ & & & \dots & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ & & & \dots & \end{pmatrix}$$

- ▶ Predict expected returns using TF-IDF
 - ▶ words \rightarrow expected returns
- ▶ Cross-sectional LASSO regression:

$$r_i = \alpha + \beta_1 \text{word}_{i,1} + \beta_2 \text{word}_{i,2} + \dots + \beta_N \text{word}_{i,N} + \epsilon_i$$

TF-IDF



→ TF-IDF =

$$\begin{pmatrix} & \text{word1} & \text{word2} & \dots & \text{word N} \\ & & \dots & & \\ & & \dots & & \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ & & \dots & & \end{pmatrix}$$

- ▶ Predict expected returns using TF-IDF
 - ▶ words → expected returns
- ▶ Cross-sectional LASSO regression:

$$r_i = \alpha + \beta_1 \text{word}_{i,1} + \beta_2 \text{word}_{i,2} + \dots + \beta_N \text{word}_{i,N} + \epsilon_i$$

- ▶ Most β s are exactly 0.

t:

t :

$t - 1$

t :

$t - 1$

oil
(1)

loan
(1)

company
(-1)

t :

$t - 1$

oil
(1)

loan
(1)

company
(-1)

$t - 2$

t :

$t - 1$	oil (1)	loan (1)	company (-1)
$t - 2$	software (0.5)	healthcare (1.1)	stock (-1.5)

t :

$t - 1$	oil (1)	loan (1)	company (-1)
$t - 2$	software (0.5)	healthcare (1.1)	stock (-1.5)
\vdots		\vdots	
$t - \tau$			

t : _____

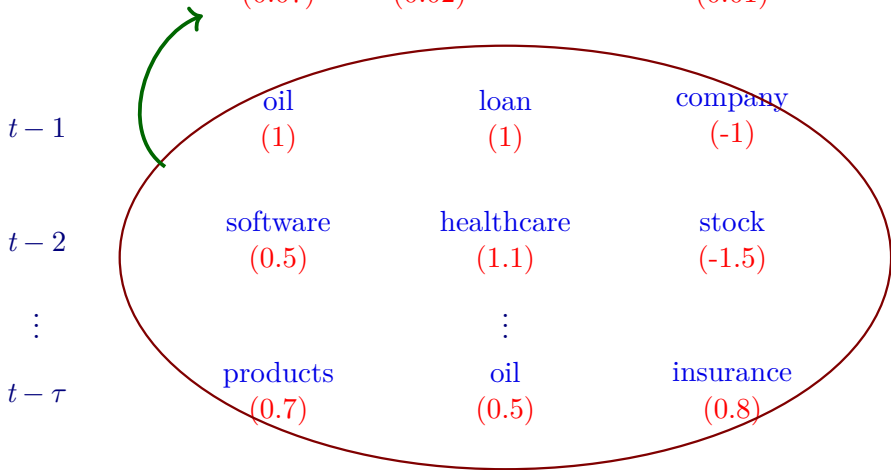
$t - 1$	oil (1)	loan (1)	company (-1)
$t - 2$	software (0.5)	healthcare (1.1)	stock (-1.5)
\vdots		\vdots	
$t - \tau$	products (0.7)	oil (0.5)	insurance (0.8)

t : _____

$t - 1$	oil (1)	loan (1)	company (-1)
$t - 2$	software (0.5)	healthcare (1.1)	stock (-1.5)
\vdots		\vdots	
$t - \tau$	products (0.7)	oil (0.5)	insurance (0.8)

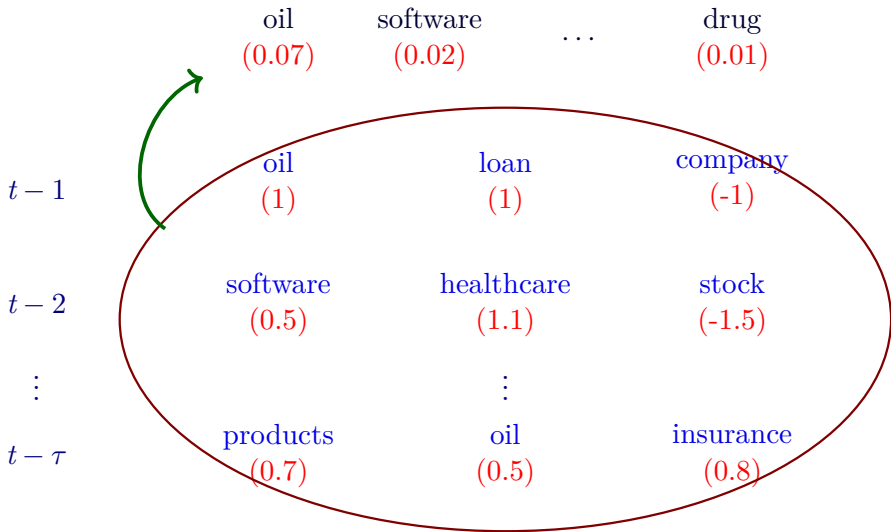
t : _____

oil (0.07) software (0.02) ... drug (0.01)



$$\tilde{r}_i = \beta_1 \text{word}_{i,1} + \beta_2 \text{word}_{i,2} + \cdots + \beta_N \text{word}_{i,N}$$

t : _____



Out-of-sample portfolios

	mean	α_{CAPM}	α_{FF4}	α_{FF8}
1	-0.33 [-0.54]	-1.47*** [-3.58]	-1.06*** [-3.61]	-0.85*** [-3.12]
2	0.33 [0.72]	-0.65** [-2.57]	-0.39** [-2.40]	-0.20 [-1.33]
3	0.53 [1.23]	-0.43** [-2.09]	-0.21* [-1.77]	-0.08 [-0.67]
4	0.70* [1.73]	-0.23 [-1.36]	-0.05 [-0.50]	0.06 [0.63]
5	0.74* [1.88]	-0.18 [-1.14]	-0.01 [-0.11]	0.07 [0.73]
6	0.85** [2.22]	-0.04 [-0.27]	0.13 [1.52]	0.17** [2.02]
7	1.01** [2.60]	0.11 [0.66]	0.25** [2.36]	0.29*** [2.69]
8	0.97** [2.59]	0.11 [0.67]	0.24** [2.36]	0.27** [2.54]
9	1.01*** [2.76]	0.17 [1.05]	0.29** [2.57]	0.33*** [2.99]
10	1.23*** [2.83]	0.36 [1.32]	0.43** [2.03]	0.51** [2.43]
LS	1.56*** [3.41]	1.83*** [4.05]	1.49*** [4.25]	1.37*** [4.16]

Out-of-sample portfolios (excluding tiny stocks)

	mean	α_{CAPM}	α_{FF4}	α_{FF8}
1	0.24 [0.43]	-0.93*** [-2.97]	-0.58*** [-2.92]	-0.46** [-2.46]
2	0.70* [1.67]	-0.28 [-1.61]	-0.08 [-0.77]	-0.01 [-0.06]
3	0.70* [1.79]	-0.24* [-1.71]	-0.09 [-1.02]	-0.07 [-0.89]
4	0.77** [2.04]	-0.14 [-1.08]	0.05 [0.05]	0.04 [0.64]
5	0.92** [2.50]	0.04 [0.30]	0.15** [1.99]	0.16** [2.06]
6	0.83** [2.25]	-0.06 [-0.46]	0.05 [0.69]	0.03 [0.39]
7	0.90** [2.55]	0.05 [0.43]	0.14** [2.07]	0.12* [1.72]
8	0.95*** [2.72]	0.11 [0.94]	0.19** [2.52]	0.16** [2.13]
9	0.80** [2.33]	-0.01 [-0.10]	0.05 [0.59]	0.02 [0.28]
10	1.00** [2.51]	0.18 [0.76]	0.17 [0.88]	0.15 [0.76]
LS	0.77* [1.78]	1.11*** [2.68]	0.75** [2.44]	0.60** [2.11]

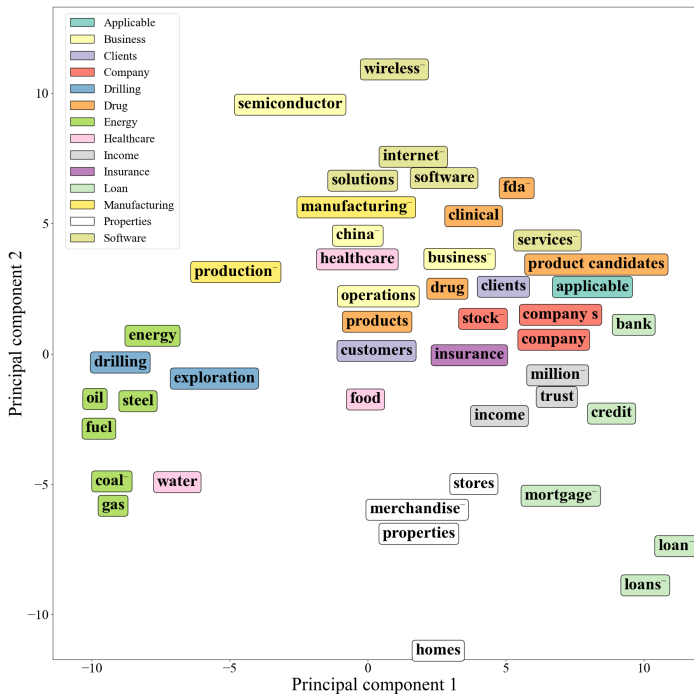
What are the risky words?

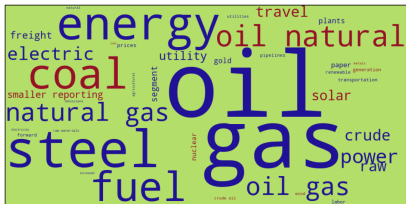
What are the risky words?

- ▶ 46 main risky words (appear more than 60 times)

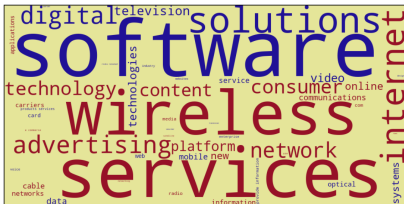
What are the risky words?

- ▶ 46 main risky words (appear more than 60 times)
- ▶ Grouped into 14 clusters
 - ▶ Using Embeddings

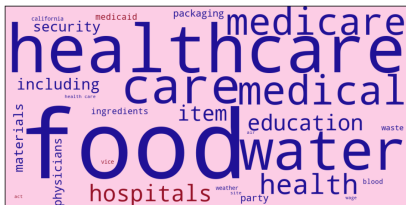




(a) Energy



(b) Software



(c) Healthcare



(d) Company

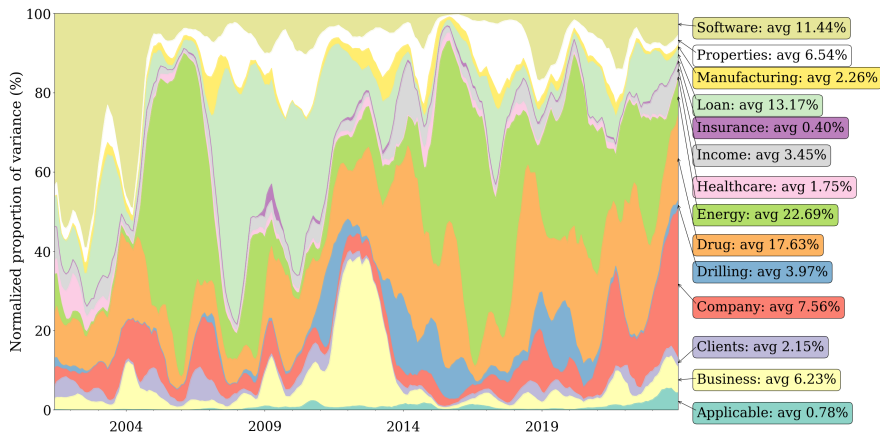
Figure: Word clouds of different clusters

Decomposing variance of expected returns

$$\tilde{r}_{i,t} = \underbrace{\sum_{\text{word}_j \in \text{energy}} b_{j,t} \text{word}_{j,i,t}}_{\text{cluster energy}} + \underbrace{\sum_{\text{word}_j \in \text{drug}} b_{j,t} \text{word}_{j,i,t}}_{\text{cluster drug}} + \cdots + \underbrace{\sum_{\text{word}_j \in \text{company}} b_{j,t} \text{word}_{j,i,t}}_{\text{cluster company}}.$$

Decomposing variance of expected returns

$$\tilde{r}_{i,t} = \underbrace{\sum_{\text{word}_j \in \text{Energy}} b_{j,t} \text{word}_{j,i,t}}_{\text{cluster energy}} + \underbrace{\sum_{\text{word}_j \in \text{drug}} b_{j,t} \text{word}_{j,i,t} + \dots}_{\text{cluster drug}} + \underbrace{\sum_{\text{word}_j \in \text{company}} b_{j,t} \text{word}_{j,i,t}}_{\text{cluster company}}$$



Predicting returns using each risky cluster

	mean	α_{CAPM}	α_{FF4}	α_{FF8}
drilling	1.04***	1.35***	1.23***	1.05***
energy	0.91***	1.13***	0.96***	0.99***
loan	0.78***	0.83***	0.70***	0.64***
properties	0.55**	0.58**	0.51**	0.40*
clients	0.54***	0.55***	0.48***	0.36**
manufacturing	0.53*	0.78***	0.67**	0.57**
company	0.52***	0.66***	0.59***	0.53***
income	0.39*	0.46**	0.41**	0.35*
drug	0.37	0.44	0.35	0.35
insurance	0.30	0.64**	0.58**	0.42
software	0.29	0.42**	0.33**	0.22
business	0.28	0.51**	0.38*	0.19
healthcare	0.25*	0.25*	0.23	0.16
applicable	0.07	0.13	0.10	-0.02

Additional Results

Additional Results

- ▶ Risky words and industries, characteristics, sentiments, and text features

Additional Results

- ▶ Risky words and industries, characteristics, sentiments, and text features
- ▶ Data mining and overfitting concerns?

Additional Results

- ▶ Risky words and industries, characteristics, sentiments, and text features
- ▶ Data mining and overfitting concerns?
- ▶ Risky words and economic uncertainty

Additional Results

- ▶ Risky words and industries, characteristics, sentiments, and text features
- ▶ Data mining and overfitting concerns?
- ▶ Risky words and economic uncertainty
- ▶ A factor of Risky Words

Literature

1. Studying risk sections:

- ▶ Campbell et al. (2014), Lopez-Lira (2023), Bao and Datta (2014)

Literature

1. Studying risk sections:

- ▶ Campbell et al. (2014), Lopez-Lira (2023), Bao and Datta (2014)

2. Measuring specific types of risks across firms:

- ▶ Florackis et al. (2023), Hanley and Hoberg (2019), Tarek A Hassan et al. (2019), Tarek Alexander Hassan et al. (2023), Tarek A Hassan et al. (2023), Kim, Muhn, and Nikolaev (2023)

Literature

1. Studying risk sections:

- ▶ Campbell et al. (2014), Lopez-Lira (2023), Bao and Datta (2014)

2. Measuring specific types of risks across firms:

- ▶ Florackis et al. (2023), Hanley and Hoberg (2019), Tarek A Hassan et al. (2019), Tarek Alexander Hassan et al. (2023), Tarek A Hassan et al. (2023), Kim, Muhn, and Nikolaev (2023)

3. Textual analysis in Asset Pricing:

- ▶ Loughran and McDonald (2011), García, Hu, and Rohrer (2023), Bybee et al. (2024), Bybee, Kelly, and Su (2023), Hoberg and Phillips (2016), Cohen, Malloy, and Nguyen (2020), Cao et al. (2023), Ross et al. (2024)



Risky Words and Returns*

Sina Seyfi[†]

December, 2024

ABSTRACT

To discover dynamic risks that determine the expected stock returns, I develop a method to predict returns through the text of firms' risk disclosures. By cross-sectionally regressing returns on the text of risk disclosures, I find certain words in the risk discussions (defined as "risky words") that have independent predictive power for the cross-section of stock returns: an out-of-sample strategy that times "risky words" earns up to 22% annual alpha between 2005-2023. Then I group risky words into 14 orthogonal clusters that are jointly eminent determinants of expected returns. Firm characteristics, industries, sentiments, and previously discovered text features do not explain the results.

Keywords: Risk sections, stock returns, textual analysis, asset pricing, machine learning

*I acknowledge financial support from OP Group Research Foundation.

[†]PhD student of Finance, Aalto University, School of Business, Department of Finance. sina.seyfi@aalto.fi, sinaseyfi.com

I. Introduction

Some stocks offer higher expected returns because they are riskier. Thus, exposure to the sources of systematic risk are true predictor of expected stock returns. Identifying the true sources of systematic risks, however, is not a straightforward task. First, there exist quite a few sources of risk that a firm potentially might be exposed to.¹ Second, even after identifying all potential sources of risks, it remains challenging to simultaneously measure the firm-level exposure to all of these risks. Third and more importantly, we are interested in understanding which sources of risks have *independent* predictive power for the cross-section of expected stock returns, that is, investors demand a higher expected return for bearing those risks, after controlling for all other sources of risk.

In this paper, I propose an intuitive and interpretable method to predict expected stock returns using all available sources of risks that firms disclose in their annual reports. The firm-level text of risk disclosures contains granular information that elaborates on the most important risks a firm faces.² In an innovative way, I use the risk section of 10-K annual reports as the predictor of future expected returns, which has a strong economic motivation: if expected returns are a function of firms' time-varying³ exposure to the sources of risks, the risk disclosure of a firm should determine the future expected returns. My objective is to reveal "which of the disclosed risks predict expected returns and to what extent". My method particularly identifies all sources of risks that investors care about, and have marginal predictive power for the cross-section of stock returns.

In a nutshell, first I find specific words in the risk sections (which I term as "risky words") that predict the cross-section of stock returns. I show that these risky words have a strong out-of-sample predictive power: a long-short equally-weighted strategy that buys (sells) firms with high (low) emphasis on risky words generates more than 22% annual market-adjusted returns between 2005-2023. When excluding tiny stocks, the risky words strategy still generates 13% annual alpha. I find that there are around 50 main risky words that constantly predict future returns between 2005-2023. Next, using a word embedding model (which I train on the text of all risk sections), I cluster these risky words so that those semantically related are clustered together. I find 14 clusters of words—including energy, healthcare, loan, and manufacturing—that are *jointly* the most predictive of future returns.

How do I determine risky words that are associated with expected returns in the first place? Loosely speaking, I cross-sectionally regress returns on the *text* of the risk sections. More formally, I begin with vectorizing the risk discussions using the TF-IDF⁴ method, wherein each

¹Examples of such risks include, but are not limited to, the political risk (Hassan, Hollander, Van Lent, and Tahoun 2019), the climate change risk (Sautner, Van Lent, Vilkov, and Zhang 2023, Li, Shan, Tang, and Yao 2024), epidemic diseases risks (Hassan, Hollander, Van Lent, Schwedeler, and Tahoun, 2023a; Davis, Hansen, and Seminario-Amez 2020), Carbon-transition risk (Bolton and Kacperczyk, 2023), environmental policy uncertainty (Hsu, Li, and Tsou, 2023), country risk (Hassan, Schreger, Schwedeler, and Tahoun, 2023b), cybersecurity risk (Florackis, Louca, Michaely, and Weber, 2023), medical innovation risk (Koijen, Philipson, and Uhlig, 2016), and Brexit risk (Hassan, Hollander, Van Lent, and Tahoun, 2023c).

²See Lopez-Lira (2021, 2023) for a discussion around "Why do managers disclose risks accurately?" See Campbell, Chen, Dhaliwal, Lu, and Steele (2014) for empirical evidence.

³I assume a dynamic-risk model, wherein the exposures to each risk varies over time. Prior literature such as Patton and Verardo (2012) and Savor and Wilson (2016) advocate this time-varying risk model.

⁴TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a method to measure how important a

risk section is represented as an ultra-high dimensional vector. Every single word in the risk section (such as *climate*) and also each bigram (such as *climate change*) represent a unique dimension, where the higher values indicate a higher importance of the words (or bigrams) in the document.⁵ Then I run cross-sectional LASSO⁶ regressions of returns on the TF-IDF vector of the most recent risk disclosure. The LASSO regression, augmented by cross-validation to shield against over-fitting and noise, sets most of the coefficients exactly to zero, and only a small set of risky words that predict returns end up with a non-zero coefficient. The LASSO regression thus takes the form of:

$$r_{i,t+1} = a_t + b_{1,t}\text{word}_{i,1,t} + b_{2,t}\text{word}_{i,2,t} + \cdots + b_{n_t,t}\text{word}_{i,n_t,t} + \epsilon_{i,t} \quad (1)$$

where $r_{i,t+1}$ is the return of firm i at the end of time $t + 1$, and $\text{word}_{i,j,t}$ is the corresponding value of word j in the risk section of firm i at time t . If firm i has not mentioned this word in time t , then $\text{word}_{i,j,t} = 0$. The higher value for $\text{word}_{i,j,t}$ indicates a higher emphasis of firm i on word j at time t . a_t is the intercept and n_t is the number of all words that appear in all of the available risk sections at time t .

To predict expected returns at each time t , I run the cross-sectional LASSO regressions of returns on the vectors of risk sections in the past and find the coefficients of risky words, b_j s. Then, using the time-series average of these coefficients, I predict stock returns with the most recent risk disclosures at time t . This methodology is similar to the traditional Fama-MacBeth regressions where predictors are the risk sections. The intuition behind this method is simple: if the usage of certain words in the risk sections has been associated with higher returns in the past, then these words are most likely indicators of systematic risks that investors demand a premium for bearing those risks. The higher emphasis of some firms on these words also indicates the higher firm exposure to these risks. Therefore, these words should determine the future expected returns as well, and firms that use these words today will likely have higher expected returns in the future. This method allows us to measure the marginal effect of each risky word on the future returns conditioned on all other sources of disclosed risks.

My method at the same time *tames* and *times* a large set of risky words. As an example, consider the risky word *oil*.⁷ A positive coefficient on the word *oil* means that the investors are compensated with higher returns when holding stocks exposed to the *oil* risk, taking into account all other sources of risks. However, when the oil price drops significantly between 2014-2016, the coefficient of the word *oil* turns negative in this period—indicating a bad state of economy for firms exposed to *oil* risk. Thus, the risky words' coefficients serve as a proxy for estimating the state of the economy. I find that the risky words' coefficients are highly time-varying and follow strong cyclical patterns. Consequently, instead of holding assets that are merely exposed to different risks, the risky words strategy simultaneously *times* different

word is to a document, considering both how often it appears in that document and how common it is overall.

⁵I loosely use the word "word" for both unigrams and bigrams.

⁶Least Absolute Shrinkage and Selection Operator

⁷The risky word *oil* is the most repeated word with a non-zero coefficient in the cross-sectional regressions. Out of 288 monthly regressions of returns on the risk sections between 2000-2023, the word *oil* has a non-zero coefficient 173 times. The next most repeated words with non-zero coefficients are *clinical* (162 times), *products* (161), *gas* (152), *company* (142), and *loans* (139).

sources of risks based on the returns reactions.⁸

Importantly, the positive or negative association of risky words to the expected returns is not related to the sentiment (whether measured based on Loughran and McDonald (2011) dictionary or Machine Learning dictionary of García, Hu, and Rohrer (2023)). I show that, while the majority of risk sections have negative sentiments, most of the risky words positively predict expected returns—consistent with the risk-based explanation.

Not only sentiments but also industries fail to explain the results. I show that firms in all of the sectors are exposed to all of the risky words simultaneously. For instance, firms in the health sector are to some degree exposed to the oil and other energy-related words, and vice versa. In fact, only 25% of shares of the word oil and other energy-related terms in the risk disclosures belong to firms in the energy industry, and the rest comes from other sectors. On the other hand, firms in the energy sector allocate only 18% of their discussions to energy-related words. The industry-adjusted returns of risky words strategy still generate 0.84% ($t = 3.45$) CAPM monthly alpha between 2005-2023, and the predictive power of risky words does not diminish when controlling for industry fixed effects.

Additionally, Li's (2008) readability of risk sections, length of the risk sections, changes in the risk section (Cohen, Malloy, and Nguyen, 2020), and firm characteristics do not explain the performance of risky words strategy. Nor this strategy is explained by common factor models of Carhart (1997), Fama and French (2015) five factors augmented by momentum, short-term and long-term reversal factors, and a 14-factor model provided by Jensen, Kelly, and Pedersen (2021).⁹ On the contrary, Jensen et al. (2021) factors barely generate significant alphas when regressed on the long-short portfolio of risky words strategy. The risky words portfolio explains these factors better than even the market factor: the Gibbons, Ross, and Shanken (1989, GRS) F -statistic—testing whether all these factors jointly have a zero alpha—drops from 4.50 (when we use the market portfolio as the regressor) to 3.98 (when the risky words portfolio is the regressor). Especially, the momentum strategy generates a negative alpha when regressed on the long-short portfolio of risky words strategy.

Around 1000 unique risky words have non-zero coefficients between 2005-2023. However, some of these words have the same semantic meaning and refer to the same concepts. For example, the words stockholders and shareholders, or costs and expenses are used interchangeably and refer to the same source of risk. To bring some structure to the “zoo” of risky words, I employ a “word embedding” (word2vec) technique to find and cluster words with similar semantic meanings. Especially, I train a word2vec model on the text of all risk sections. The idea is that every word is represented as a high-dimensional vector and words with same semantic meanings share similar vector representation.¹⁰

⁸This approach is similar to Neuhiel, Randl, Reschenhofer, and Zechner (2023) who show that “Timing the factor zoo” would significantly improve the performance of investment strategies.

⁹Jensen et al. (2021) introduce a factor taxonomy and group more than 150 factors in 13 groups based on the correlations. These 13 factors include accruals, debt issuance, investment, low leverage, low risk, momentum, profit growth, profitability, quality, seasonality, short-term reversal, size, and value. I augment these 13 factors with the market portfolio and use them as a comprehensive factor model for asset pricing tests.

¹⁰The word embedding technique is a groundbreaking method in natural language processing (NLP), and they are proven to be exceptionally powerful at capturing the semantic meaning between words. I verify that the vectors obtained by training the model on the text of risk sections highly capture the semantic meanings between words. For instance, the most similar vector to the vector of the word stock is for the word shares. Or, the most similar

Using the vector representation of risky words, I then cluster these risky words such that words with similar vectors are grouped together. By grouping words with similar semantic meanings (measured as how likely they appear in similar contexts in the risk sections), I find sources of systematic risks and risky words related to them. Using the coefficients obtained from LASSO regressions, I further measure to what extent each risk predicts expected returns. I finally decompose the cross-sectional variance in expected returns to the variance of sources of risks. Risks related to **energy**, **drug** and **loan** have the highest contribution in explaining the cross-sectional variance of stock returns.

Each of the risky words belongs to one of the following clusters: applicable, business, clients, company, drug, drilling, energy, healthcare, income, insurance, loan, manufacturing, properties, and software. I show that words related to each cluster exclusively predict future returns, but the predictability is highest when these words are jointly used for the prediction. That a wide range of topics is associated with expected returns shows that the sources of systematic risks are quite high-dimensional, consistent with the emerging literature that the asset prices are not explained by a small set of factors.¹¹

Because the risk sections are quite lengthy and inherently noisy, an important concern is the potential for data mining or overfitting on noise when regressing returns on thousands of predictors (i.e., words in the risk disclosures). To address these concerns, I use LASSO regression combined with 5-fold cross-validation in each cross-sectional analysis, which serves as a strong safeguard against noise. I validate the model's robustness by observing that, in the cross-sectional regressions, the model consistently picks the same set of around 50 significant words (from thousands of possibilities) over time, indicating that these words are not chosen by chance. If the results were due to data mining or overfitting on noise, then it would be very plausible to see the model selecting more variables as the length of the risk disclosures increases, as it provides more opportunity for spurious correlations between words and returns. The fact that this does not happen and that the number of significant predictors remains relatively consistent and irrelevant to the number of all of the words supports the idea that the findings are not merely a result of data mining or overfitting. Additionally, when randomly reassigning risk sections to firms, the predictive ability of the model entirely vanishes—no significant risky words emerge, supporting the specificity of the results to actual firm disclosures. Moreover, by simulating random matrices in place of the original risk sections, I confirm that the cross-validated LASSO does not select random noise as predictors. Finally, I assess the performance of the “risky words”

vectors to the vector of the word **ESG** is for the words **sustainability**, **governance** and **climate**. If $\vec{x}(W)$ shows the vector representation of a given word **W**, I find that the

$$\vec{x}(\text{acquisition}) - \vec{x}(\text{buy}) + \vec{x}(\text{sell}) \approx \vec{x}(\text{divestiture}) \quad \text{or} \quad \vec{x}(\text{sale}).$$

That is, if we subtract the vector representation of the word **buy** from the vector of the word **acquisition** and add the vector of **sell**, the resulting vector is most similar to the vectors of **divestiture** or **sale**. Another example is,

$$\vec{x}(\text{inflation}) - \vec{x}(\text{increase}) + \vec{x}(\text{decrease}) \approx \vec{x}(\text{deflation}).$$

Vector representation of risky words has other implications such as mapping the unstructured text of risk sections (or other sources of financial text) to the firm value, and other economic indices. [Hoberg and Phillips \(2024\)](#) use a Doc2vec approach to study the scope of operation. Another early attempt at using word embedding idea, but on investor holdings rather than actual words, is asset embedding of [Gabaix, Koijen, Richmond, and Yogo \(2023\)](#).

¹¹For the discussion around high-dimensionality of risk factors, see [Bryzgalova, Huang, and Julliard \(2022\)](#), [Kozak, Nagel, and Santosh \(2020\)](#), [Giannone, Lenza, and Primiceri \(2021\)](#), [Jensen et al. \(2021\)](#), and [Seyfi \(2022, 2023, 2024\)](#).

strategy exclusively out-of-sample, validating the reliability and generalizability of the picked words.

At the aggregate level, risky words predict economic uncertainty. Notably, I find that more risky words get a non-zero coefficient is an indication that the expected returns are affected by a large number of systematic risks, and thus, economic uncertainty is high in the following months. Conversely, in the states of economics with low uncertainty, there are less risky words that affect stock prices. I show that there is a significant correlation between the aggregate number of risky words and future VIX (30-day option-implied volatility in the S&P500 index), and the economic policy uncertainty index of [Baker, Bloom, and Davis \(2016\)](#).

II. Literature

My paper is related to the flourishing literature of textual analysis in finance.¹² A growing number of papers exclusively study the risk disclosure of firms' reports. [Campbell et al. \(2014\)](#) provide first shreds of evidence that firm-specific risk disclosures are reflected in the systematic and idiosyncratic risks, and contain useful information for the investors. [Gaulin \(2017\)](#) rules out the concerns of purely "copy and paste" disclosure in the risk sections. [Bao and Datta \(2014\)](#) and [Lopez-Lira \(2023\)](#) apply a topic modeling on the risk section and find firm exposure to each topic.

Another set of papers focuses on text data to measure specific types of risks across firms. For instance, [Florackis et al. \(2023\)](#) use risk sections to measure and quantify the cybersecurity risk. [Hanley and Hoberg \(2019\)](#) find emerging risks in the financial sectors. [Hassan et al. \(2019\)](#) use earning calls to measure the political risk. [Hassan et al. \(2023c\)](#) measure the global impact of Brexit risk, and [Hassan et al. \(2023a\)](#) construct text-based measures of exposure to epidemic diseases, such as Covid-19. Recently, [Kim, Muhn, and Nikolaev \(2023\)](#) use chatGPT to study the firm-level measures of political, climate, and AI-related risks.

My paper, on the other hand, deviates from the literature as my approach is a supervised method that searches for risky words that *jointly* predict the cross-section of stock returns, instead of identifying some or all of the risks that are mentioned by the firms. My approach times different sources of risks based on the recent price reactions, and sheds light on the time-varying relationship between a set of disclosed risk factors and returns.

Another branch of the literature utilizes textual analysis methods in asset pricing. [Loughran and McDonald \(2011\)](#) and [García et al. \(2023\)](#) focus on measuring the sentiment of Finance words. [Bybee, Kelly, Manela, and Xiu \(2024\)](#) measure the state of the economy using the text of newspapers, and [Bybee, Kelly, and Su \(2023\)](#) create a "narrative" factor from the text of the Wall Street Journal. [Hoberg and Phillips \(2016\)](#) define new industries based on the similarity of business descriptions from 10-K filings. [Cohen et al. \(2020\)](#) show that changes to 10-Ks are associated with lower future performance. [Cao, Green, Lei, and Zhang \(2023\)](#) study the text of Expert Network Calls, and find that the tone of expert calls predicts future returns.

My paper is related to the asset pricing literature by addressing the question of what are the sources of risks determining the expected returns. Examples of these papers include [Fama and](#)

¹²See [Loughran and McDonald \(2016, 2020\)](#) for a comprehensive review.

French (1992, 1993, 2008, 2015, 2016, 2018), Feng, Giglio, and Xiu (2020), Harvey and Liu (2021), and Kozak, Nagel, and Santosh (2018), among others. Unlike these papers, I opt for a textual analysis approach to discover the sources of systematic risks affecting expected returns.

III. Data and Summary Statistics

In this section, I introduce the data and present the summary statistics.

A. Risk Sections

All the firms' 10-K filings are publicly available on the SEC's EDGAR website. However, I download the text of the annual reports (10-K, 10-K-A, 10-K405, 10-KT, 10-K405-A, and 10-KT-A) from Loughran and McDonald website, where they provide raw and cleaned versions of firm reports.¹³ I obtain the company name, CIK code,¹⁴ SIC code, and the filing date from the reports itself. I opt for the following algorithm to extract the risk sections from 10-K filings and its variants. Normally, the risk sections come under the "Item 1A - risk factors" and end before either "Item 1B" or "Item 2". Therefore, I find all the patterns that capture various formats of "item 1A - Risk Factors" until "Item 1B" or "Item 2".¹⁵ One caveat is that as the risk sections do not come as separate downloadable files, there might be some measurement errors in extracting the risk sections. However, I manually check a random subset of risk sections and verify that the extraction has a high accuracy. I drop all observations that a risk factor is not found. This leads to finding around 108,000 unique risk sections between 2000 and 2023.

For most firms, the fiscal year matches the end of the calendar year. Thus, most companies file their reports in February and March.¹⁶ However, around one-third of firms file their reports in other months. For this reason, I perform the analysis at a monthly level to ensure that all available information is used when constructing the risky words portfolios. Figure A1 in the Appendix shows the average of files (that contained a risk section) from 1995-2023. Figure A2 in the Appendix, shows the number of files in each year (panel a), and the average number of files per day (panel b).

Since 2005, the SEC has required publicly traded companies to include a "Risk Factors" section in their Form 10-K filings. Since then, the length of risk sections monotonically increases. For instance, in 2006, the average length of the "Risk Factors" sections was approximately 5,000 words. By 2023, the average length of these sections has tripled to around 16,000 words. Figure A3 in the Appendix shows the average number of words in the risk sections per year.

¹³I thank Loughran and McDonald for providing data in their website. The data is available for download here: <https://sraf.nd.edu/sec-edgar-data/cleaned-10x-files>.

¹⁴The CIK (Central Index Key) code is a unique identifier assigned by the Securities and Exchange Commission (SEC) to public companies and certain other entities that file disclosure documents with the SEC.

¹⁵By various formats of "Item 1A - Risk Factors", I refer to cases such that "item 1A / Risk Factors" or "item-1A. risk factors". If there are multiple of such cases, I choose the latest one. If "item 1A" is not found (which is the case in many firms' reports), I find only "risk factors" until "item 1B" or "item 2".

¹⁶Based on the filing deadlines and requirements for Form 10-K in the SEC's rules under the Securities Exchange Act of 1934, specifically in Regulation S-K, companies classified as large accelerated filers, with a public float of \$700 million or more, must file their 10-K within 60 days after the end of their fiscal year. Accelerated filers, those with a public float between \$75 million and \$700 million, have 75 days to file. Non-accelerated filers, with a public float less than \$75 million, are given 90 days to file their 10-K reports.

B. Other Data

I obtain monthly stock returns for all firms listed on the NYSE, AMEX, and NASDAQ from 2005 through the end of 2023, sourced from CRSP (Center for Research in Security Prices). Then, using the CIK code obtained from Compustat, I match the risk sections to the monthly returns. For each firm at each month, I assign the most recent risk section. For instance, if a firm previous risk section is filed on 15.2.2020 and the next risk section is filed on 20.3.2021, I use the former risk section for all the months between 3.2020 and 3.2021, inclusively. This way, I end up with more than 1.23 million firm-month observations from 2000 to the end of 2023.

I obtain the risk-free rate by using the one-month Treasury bill rate available from the Fama and French Data Library. Likewise, the excess market returns and the size, value, and momentum factors are also sourced from this library. Additionally, [Jensen et al. \(2021\)](#) introduce a classification for factors into 13 distinct categories, designed to ensure low return correlation between groups while maintaining high correlation within them. These categories encompass investment, profitability, accruals, debt issuance, low leverage, low risk, profit growth, quality, seasonality, and short-term reversal, along with size, value, and momentum. Except for the size, value, and momentum factors, which are obtained from the Fama-French library, I source the remaining factors from [Jensen et al. \(2021\)](#)'s global factor database.

For firm characteristics, I obtain beta, illiquidity, coskewness, idiosyncratic volatility, firm age, maximum daily returns, R&D, and asset growth from data library provided by [Chen and Zimmermann \(2021\)](#), augmented with size, momentum, short- and long-term reversal calculated from CRSP data. To address the issue of missing values in the characteristics, I adopt the methodology outlined by [Gu, Kelly, and Xiu \(2020\)](#), which involves imputing missing values using their respective medians. Furthermore, I normalize all characteristics across cross-sections to achieve a mean of zero and a variance of one for each month.

To measure the sentiments, I use the Machine Learning Dictionary of Finance words based on the recent work of [García et al. \(2023\)](#). However, all the results remain the same when using the traditional dictionary of [Loughran and McDonald \(2011\)](#). The overall risk sections' sentiments are negative. Figure 1 shows the average number of positive and negative words in the risk sections based on the [García et al. \(2023\)](#) dictionary, with the negative words outnumbering the positive words, on average in each year. To measure the sentiment of each document, I use the sentiment score, defined as

$$\text{Sentiment Score} = \frac{\text{count of positive words} - \text{count of negative words}}{\text{count of positive words} + \text{count of negative words}}.$$

The sentiment score ranges from -1 to +1, where -1 indicates a completely negative sentiment, +1 indicates a completely positive sentiment, and 0 indicates neutrality. The corresponding figure based on the traditional dictionary of [Loughran and McDonald \(2011\)](#) is shown in Figure A5 in the Appendix. The average sentiment score of the risk sections between 2000 and 2023 is -0.33 based on the [Loughran and McDonald \(2011\)](#) dictionary. If using the Machine Learning dictionary of [García et al. \(2023\)](#), the average sentiment score for the same time period is -0.19, both numbers significantly different from 0.

Lastly, I also obtain the VIX volatility index from Yahoo! Finance, and the economic uncer-

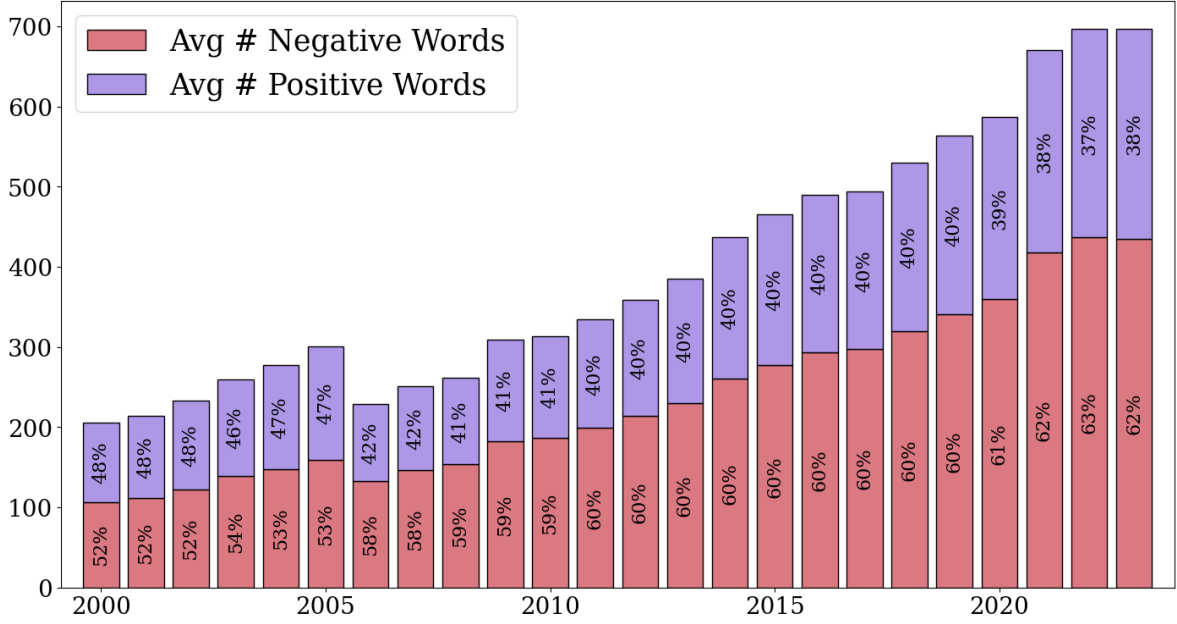


Figure 1. The average number of positive and negative words in risk sections per year

This Figure illustrates the annual average counts of positive (blue) versus negative (red) words within the risk sections, consistently revealing a higher frequency of negative words each year. The positive and negative words are calculated based on the Machine Learning dictionary of [García et al. \(2023\)](#). The corresponding figure based on the traditional dictionary of [Loughran and McDonald \(2011\)](#) is shown in figure A5.

tainty index based on the work of [Baker et al. \(2016\)](#).

C. Vectorizing the Risk Sections

The first objective is to discover which risks (or risky words) predict expected returns. Thus, I start by converting each risk section to an ultra-high-dimensional vector and use this vector as the predictor of expected returns. Especially, for each month, I create the TF-IDF (Term Frequency–Inverse Document Frequency) matrix based on the most recent risk disclosures and show that this matrix predicts future expected returns.

TF-IDF method is a common statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance of a word increases proportionally to the number of times it appears in a document but is offset by the frequency of the word in the corpus. This helps to adjust for the fact that some words appear more frequently in general. Moreover, I normalize each row in the document so each vector has a l_2 norm of 1, regardless of the length of the document.

Term Frequency measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear more times in long documents than in shorter ones. Thus, I divide the term frequency by the document length (the total number of terms in the document) as a way of normalization:

$$TF(W, d) = \frac{\text{Number of times term } W \text{ appears in a document } d}{\text{Total number of terms in the document } d}$$

While computing TF, all terms are considered equally important. However, certain frequent words may appear way more times but carry little information. Thus I weigh down the frequent terms while scaling up the rare ones, by computing the following:

$$IDF(W, D) = \log \left(\frac{\text{Total number of documents } D}{\text{Number of documents with term } W \text{ in it}} \right)$$

TF-IDF is simply the TF multiplied by IDF:

$$TF.IDF(W, d, D) = TF(W, d) \times IDF(W, D)$$

This results in a weight for each word in every document. Words that occur frequently in a document but not across many documents in the corpus will have a high TF-IDF score. This indicates that these words are not only frequent in the specific document but also provide a significant amount of information about the content of that document, distinguishing it from others in the corpus.

Example case. Suppose we have three risk sections for three firms in a given month as follows:

Firm 1: Covid hurts.

Firm 2: Carbon emission and technology matter.

Firm 3: Climate change and Covid affect us.

The TF-IDF matrix of this simple example can be represented as follows:

	Covid	hurts	carbon	emission	technology	matter	climate	change	affect	us	and
Firm 1	0.34	0.94	0	0	0	0	0	0	0	0	0
Firm 2	0	0	0.49	0.49	0.49	0.49	0	0	0	0	0.18
Firm 3	0.19	0	0	0	0	0	0.48	0.48	0.48	0.48	0.19

The matrix above shows that, for instance in firm 1, only the words "Covid" and "hurts" have non-zero weights. Moreover, the word "Covid" is more important in firm 1 than firm 3, as firm 3 mentions other topics in addition to the Covid.

When creating the TF-IDF matrix of all of the risk sections, I consider both single words (unigrams) and two-word combinations (bigrams) as individual features. Furthermore, I filter out common English stop words (words such as "is", "and", "the", etc.). These words do not carry information and dropping them reduces the noise. Moreover, I drop any non-alphabetic character (such as numerics). I filter out terms that appear too infrequently (in less than 5% of the documents) or too frequently (in more than 95% of the documents) in the corpus. Terms ubiquitous across more than 95% likely derive from standard templates and do not predict expected returns. Conversely, terms rare in the dataset, appearing in only a few documents, typically represent idiosyncratic words unrelated to expected returns.

These filters lead to the creation of ultra-high-dimensional TF-IDF matrices as predictor of expected returns. The dimension of the TF-IDF matrix exceeds 20,000 at the end of the sample with the majority of columns containing bigrams. At the beginning of the sample, the TF-IDF

matrix has around 5,000 elements. Figure A4 in the Appendix shows the average number of unique unigrams and bigrams in each year.

IV. Risky Words and Expected Returns

In this section, I elaborate on how I find the risky words at each month, and how I create a risky words strategy.

A. A Dynamic Risk and Return Model

I consider the following model

$$\mathbb{E}(r_{i,t+1}) = \beta_{1,t}\text{Risk}_{i,1,t} + \beta_{2,t}\text{Risk}_{i,2,t} + \cdots + \beta_{n,t}\text{Risk}_{i,n,t}. \quad (2)$$

where $\beta_{j,t}$ shows how much expected returns are sensitive to the j -th source of risk at time t , and $\text{Risk}_{i,j,t}$ shows how much firm i is exposed to the Risk_j at time t . Without any prior assumptions, we do not know the number of risks affecting returns (n_t), the values of coefficients β_j s (which indicate the extent to which each risk predicts returns), or the exposure of each firm to each risk ($\text{Risk}_{i,j}$) at each time t . Using firm-level text from the risk sections, I estimate these unknowns.

B. Finding Risky Words

To find out which words in the risk section predict expected returns at each month t , I examine the return predictability in the following form:

$$r_{i,t+1} = a_t + b_{1,t}\text{word}_{i,1,t} + b_{2,t}\text{word}_{i,2,t} + \cdots + b_{n_t,t}\text{word}_{i,n_t,t} + \epsilon_{i,t} \quad (3)$$

In the above expression, $\text{word}_{i,j,t}$ shows the weight of j -th word for firm i at time t . Coefficient $b_{j,t}$ shows how much word j at time t has predictive power for the returns at time $t + 1$, and a_t is the intercept. The number of words is n_t , $\epsilon_{i,t}$ are the error terms. $\text{word}_{i,j,t}$ includes both unigrams and bigrams and hence the number of explanatory variables in the above regression is far more than the number of observations. Moreover, the majority of the coefficients are expected to be exactly zero, as many of the words do not have any predictive power for future returns. For these reasons, I consider solving the above regression using the LASSO optimization procedure. The LASSO shrinks most of the coefficients equal to zero, and a small set of words ends up with a non-zero coefficient.

To find the regularization parameter in LASSO at each month, I implement a cross-validation to ensure that the model generalizes well to unseen data. In short, I divide the data into 5 equal parts, each serving as a validation set once in a rotation, while the remaining 4 parts are used for training. This process, known as 5-fold cross-validation, is repeated for a range of different regularization parameters, and the one is chosen that offers the lowest average error. The details of the cross-validation are presented in Appendix A.

C. Out-of-sample Portfolios of Risky Words Strategy

To create a strategy based on the risky word at time $t + 1$, first I find the coefficient of each word in the past. The coefficients show to what extent each word predicts future returns. More specifically, to predict the returns at time $t + 1$, I find the b coefficients in the past 5 years (between $t - 59$ and t). Then I find the weighted average of these coefficients so that the more recent ones have a higher weight. More specifically, when averaging the coefficients from the past, I weigh the coefficient at time $t - t'$ with $e^{-\lambda t'}$, implying that the predictive power of risky words decreases over time. In my empirical analysis, I set $\lambda = 0.1$, but the model is robust to different choices of λ . Mathematically, the coefficient of word $_{j,t}$, shown as $\bar{b}_{j,t}$ is:

$$\bar{b}_{j,t} = \frac{1}{T} \sum_{t'=0}^{T-1} e^{-\lambda t'} b_{j,t-t'}, \quad (4)$$

where $b_{j,t-t'}$ is estimated based on the equation 3, $T = 60$ months, and $e^{-\lambda t'}$ is an exponential decay function to weights based on the time difference from the most recent date.¹⁷

Finally, I predict the return of stock i at time $t + 1$ as

$$\tilde{r}_{i,t+1} = \bar{b}_{1,t} \text{word}_{i,1,t+1} + \bar{b}_{2,t} \text{word}_{i,2,t+1} + \dots + \bar{b}_{n_t,t} \text{word}_{i,n_t,t+1}, \quad (5)$$

assuming that there are n_t words in all documents, and word $_{i,j,t+1}$ is j -th words weight in the most recent document for firm i .

Then I sort stocks into deciles based on the predicted values $\tilde{r}_{i,t+1}$, and create equally-weighted portfolios. Because the LASSO regression in equation 3 minimizes the errors on an equally weighted basis, it is more intuitive to create equally weighted portfolios. However, to make sure that results do not stem from the microcaps, I report the results for the case where microcaps are excluded (stocks that have a size below 20% NYSE percentile). I update portfolios monthly. I create out-of-sample portfolios from January 2005, until the end of December 2023. The first column of Table 1 shows the out-of-sample mean excess returns of decile portfolios, with the last row showing a long-short portfolio that buys portfolio 10 and sells portfolio 1. I show the t -statistics in parentheses. The mean returns fairly monotonically increase from portfolio 1 to 10. The long-short portfolio generates a monthly mean excess return of 1.56% ($t = 3.41$), which is economically large and statistically significant.

The second column of Table 1 shows the risk-adjusted returns of portfolios against CAPM. The long-short portfolio displays 1.83% ($t = 4.05$) monthly CAPM alpha. The third column shows the alphas with respect to the Fama-French-Carhart four-factor model (FF4). The fourth column shows an eight-factor model which consists of Fama and French (2015) five-factor model augmented with momentum, short-term and long-term reversal factors (FF8). With all factor models, the long-short portfolio strategy of risky words generates a highly large and

¹⁷ An exponential decay function models a process where a quantity diminishes at a rate proportional to its current value, resulting in a rapid decrease at the outset followed by a progressively slower decline over time. Such functions are pivotal in describing phenomena like radioactive decay, population decline, and thermal dissipation, where the pace of change decreases as the quantity lessens, reflecting a natural, gradual tapering-off characteristic of these processes. I assume the predictive power of risky words follows a similar pattern. Finding a simple time-series average also presents similar results, although the most recent coefficients show stronger predictive powers.

	All Stocks				All But Tiny Stocks			
	mean	α_{CAPM}	α_{FF4}	α_{FF8}	mean	α_{CAPM}	α_{FF4}	α_{FF8}
1	-0.33 [-0.54]	-1.47*** [-3.58]	-1.06*** [-3.61]	-0.85*** [-3.12]	0.24 [0.43]	-0.93*** [-2.97]	-0.58*** [-2.92]	-0.46** [-2.46]
2	0.33 [0.72]	-0.65** [-2.57]	-0.39** [-2.40]	-0.20 [-1.33]	0.70* [1.67]	-0.28 [-1.61]	-0.08 [-0.77]	-0.01 [-0.06]
3	0.53 [1.23]	-0.43** [-2.09]	-0.21* [-1.77]	-0.08 [-0.67]	0.70* [1.79]	-0.24* [-1.71]	-0.09 [-1.02]	-0.07 [-0.89]
4	0.70* [1.73]	-0.23 [-1.36]	-0.05 [-0.50]	0.06 [0.63]	0.77** [2.04]	-0.14 [-1.08]	0.00 [0.05]	0.04 [0.64]
5	0.74* [1.88]	-0.18 [-1.14]	-0.01 [-0.11]	0.07 [0.73]	0.92** [2.50]	0.04 [0.30]	0.15** [1.99]	0.16** [2.06]
6	0.85** [2.22]	-0.04 [-0.27]	0.13 [1.52]	0.17** [2.02]	0.83** [2.25]	-0.06 [-0.46]	0.05 [0.69]	0.03 [0.39]
7	1.01** [2.60]	0.11 [0.66]	0.25** [2.36]	0.29*** [2.69]	0.90** [2.55]	0.05 [0.43]	0.14** [2.07]	0.12* [1.72]
8	0.97** [2.59]	0.11 [0.67]	0.24** [2.36]	0.27** [2.54]	0.95*** [2.72]	0.11 [0.94]	0.19** [2.52]	0.16** [2.13]
9	1.01*** [2.76]	0.17 [1.05]	0.29** [2.57]	0.33*** [2.99]	0.80** [2.33]	-0.01 [-0.10]	0.05 [0.59]	0.02 [0.28]
10	1.23*** [2.83]	0.36 [1.32]	0.43** [2.03]	0.51** [2.43]	1.00** [2.51]	0.18 [0.76]	0.17 [0.88]	0.15 [0.76]
LS	1.56*** [3.41]	1.83*** [4.05]	1.49*** [4.25]	1.37*** [4.16]	0.77* [1.78]	1.11*** [2.68]	0.75** [2.44]	0.60** [2.11]

Table 1- Out-of-sample performance of risky words strategy

This table presents the out-of-sample performance of risky words decile portfolios sorted based on predicted returns $\hat{r}_i, t + 1$, as defined in equation (3). Portfolios are constructed monthly from January 2005 to December 2023 and are equally weighted. I run the cross-section LASSO regressions of returns on the ultra-high-dimensional vectors of the text of risk disclosures. These regressions result in a small set of words that have predictive power for returns. Then using the weighted average coefficients of these risky words, I make a prediction for the next month's returns. The four left columns show the portfolios when including all stocks. The four right columns exclude microcaps (firms below the 20th NYSE size percentile). Column mean shows the average excess returns of each portfolio. Other columns show the alpha with respect to CAPM, the Fama-French-Carhart 4 factor model, and Fama-French 5 augmented by momentum, short and long-term reversals factors (FF8). The last row shows a long-short portfolio (LS). The t -statistics are shown in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

significant alpha—all alphas are above 1.37% per month with t -statistics above 4, higher than the proposed threshold by [Harvey, Liu, and Zhu \(2016\)](#).

The four columns in the right part of Table 1 show the counterpart results for the case where the tiny stocks are excluded. The CAPM alpha drops to 1.11% per month with $t = 2.68$. The model still generates an alpha of 0.75% ($t = 2.44$) and 0.60% ($t = 2.11$) with respect to Carhart and Fama-French eight-factor model.

In section VI, I show that these performances do not stem from the text features, industry, characteristics, sentiment, or purely chance.

V. What are the Risky Words?

Now, I present the risky words that get a non-zero coefficient when regressing returns on the text of risk disclosures. I find that 46 risky words constantly appear in the cross-sectional regressions (at least 60 times). I consider these as the main risky words and group them into 14 themes based on the similarity of their semantic meanings. Then, all other risky words will belong to any of these 14 clusters.

For clustering the risky words, I use their vector representation. I start by explaining a word2vec model which I train on the words of all the risk sections to obtain the vectors for each word.

A. A word2vec model Trained on Risk Sections

To introduce some structure to the chaotic collection of risky words, I train a word embedding model on the text of all risk sections, inspired by an advanced method from natural language processing (NLP). The idea is that words that are likely to appear in similar contexts would have similar vector representations.

To create meaningful vectors for the words, I use word2Vec to transform words into continuous vectors. I train the model using a Continuous Bag of Words (CBOW) architecture. The CBOW model predicts the current word based on its context. It operates by taking a window of words surrounding the target word and using these context words to predict the target word itself. For instance, assume that in the following sentence the word *crashed* is missing: "The stock market _____ due to economic uncertainty." If we want to predict the word "*crashed*," we would use the surrounding words like *The, stock, market, due, to, economic* and *uncertainty* as the context.

The CBOW model takes these context words and learns to predict the middle word, *crashed*. By repeating this process with many sentences in the risk sections, the model learns which words typically appear together and how they are related. For example, it would learn that *stock, market,* and *crashed* often appear in contexts discussing financial downturns or volatility. Appendix B further elaborates on this algorithm.

It is worth highlighting that training the word embedding model on all of the risk sections does not introduce any look-ahead bias in my analysis: I do not use these vectors for any return predictability or out-of-sample analysis. The application of the vectors is to just cluster the risky words that are shown to have predictive power for the cross-section of stock returns from LASSO regressions. An alternative approach would be to train the model on only a base year (as [Hoberg and Phillips \(2024\)](#) do). However, it might be the case that, as the risk sections evolve over time, some words that are absent in the risk sections in the base year never get a vector. For instance, the word *covid*—that appears to have a negative impact on returns only in 2020—is missing from firms' disclosures before the start of the COVID-19 pandemic.

A.1. Validating the word2vec model

For each word in the risk sections, I obtain a 300-dimensional vector.¹⁸ I then validate that the vectors capture the semantic relation between words. Table 2 shows a few examples of the relationship between vectors of words where $x(W)$ shows the vector for the word W . For instance, if we take the vector of the word **ESG** as $x(\text{ESG})$, the most similar vector to it is $x(\text{sustainability})$ with a cosine similarity of 0.68. Likewise, the most similar vectors to the vector of **carbon** are for the words **GHG** (greenhouse gas) and **emissions**.

We can also perform mathematical operations on vectors. For instance, we add the vectors of $x(\text{artificial})$ and $x(\text{intelligence})$, the resulting vector is most similar to the vector for the word $x(\text{AI})$. In a similar way, subtracting $x(\text{buy})$ from $x(\text{acquisition})$ and adding $x(\text{sell})$ results to a vector that is most closest to the $x(\text{divestiture})$ and $x(\text{sale})$. Similarly, I observe that

$$x(\text{loan}) - x(\text{default}) \approx x(\text{assets}) - x(\text{losses}),$$

implying that the loan without any default is analogous to assets without losses.

The relationship between words also include the geographical relationship between words that relate to the locations. For instance the average vector for words **Finland** and **Norway** is most similar to the vectors of words **Sweden** and **Denmark**.

B. Main Risky Words

Armed with the vector representation of all of the individual words in the risk section, I cluster the main risky words based on the similarity of their vectors. I use Agglomerative clustering which does not require a pre-defined number of clusters. Instead, to determine the optimal clustering, I set a distance threshold parameter to control the stopping condition of the agglomerative merging process. This unsupervised approach is advantageous when the optimal number of clusters is not known a priori.

Agglomerative clustering is a type of hierarchical clustering that builds nested clusters by successively merging or splitting them based on a defined criterion. I define the distance between two clusters as the maximum distance between any two points in the clusters. Thus, I begin with each word as an individual cluster and iteratively merge the closest pairs of clusters. The merging process halts when the cosine distance between all remaining clusters exceeds 0.95, indicating that clusters are no longer sufficiently similar (i.e., have less than 5% similarity).

The algorithm groups words into clusters where the furthest pair of points between any two merged clusters has a cosine distance less than or equal to 0.95 (which corresponds to a cosine similarity greater than or equal to 5%). This ensures that words in different clusters have a cosine distance greater than 0.95 (similarity less than 5%), implying they are almost orthogonal and unrelated. Then I label the clusters using one of the words in each cluster. These 14 clusters include applicable, business, clients, company, drug, drilling, energy, healthcare,

¹⁸The choice of 300 dimensions for word embeddings is a convention and is based on empirical findings that this dimensionality effectively captures semantic relationships between words while maintaining computational efficiency. For example, the word2Vec model trained on the Google News dataset includes 300-dimensional vectors for 3 million words and phrases, demonstrating the model's capacity to represent a vast vocabulary in a compact vector space.

Original Word(s)/vector(s)	Most Similar Word(s)/vector(s)
$x(\text{ESG})$	$x(\text{sustainability})$
$x(\text{carbon})$	$x(\text{GHG}), x(\text{emissions})$
$x(\text{stock})$	$x(\text{shares}), x(\text{securities})$
$x(\text{covid})$	$x(\text{pandemic}), x(\text{coronavirus})$
$x(\text{adversely})$	$x(\text{negatively})$
$x(\text{customer})$	$x(\text{client})$
$x(\text{assurance})$	$x(\text{guarantee})$
$x(\text{artificial}) + x(\text{intelligence})$	$x(\text{AI})$
$x(\text{acquisition}) - x(\text{buy}) + x(\text{sell})$	$x(\text{divestiture}), x(\text{sale})$
$x(\text{inflation}) - x(\text{increase}) + x(\text{decrease})$	$x(\text{deflation}), x(\text{disinflation})$
$x(\text{loan}) - x(\text{default})$	$x(\text{assets}) - x(\text{losses})$
$x(\text{telephone}) - x(\text{voice}) + x(\text{fax})$	$x(\text{telecopy})$
$x(\text{Finland}) + x(\text{Norway})$	$x(\text{Sweden}), x(\text{Denmark})$

Table 2- Validating the word embeddings

This table validates the relationships captured by the word embeddings, showcasing the semantic connections between different words based on their vector representations. By training a word2Vec model using the Continuous Bag of Words (CBOW) architecture, I generate 300-dimensional vectors that reflect the contextual similarity of words in risk sections. The right column shows the vectors of a word or the operation of some vectors corresponding to the most similar vector in the left column, where similarity is measured as the cosine of the angle between two vectors. $x(W)$ represents the vector for the word W . For instance, vector [sustainability](#) is most similar to the vector of the word [ESG](#). Similarly, vectors of words [pandemic](#) and [coronavirus](#) are most similar to the vector of [covid](#). Additionally, mathematical operations on vectors reveal logical relationships, as seen when combining [artificial](#) and [intelligence](#) to approximate [AI](#). Geographic associations are also captured, where averaging vectors for [Finland](#) and [Norway](#) yields similarities to [Sweden](#) and [Denmark](#). These validations confirm that the embeddings effectively encode meaningful relationships, useful for clustering and analyzing risk-related language.

income, insurance, loan, manufacturing, properties, and software. I plot these words and their cluster labels in Figure 2. The location of each word in Figure 2 is based on the two first principal components (PC) from their 300-dimensional vectors.

Figure 2 shows that words such as [coal](#), [energy](#), [fuel](#), [gas](#), [oil](#), and [steel](#) are grouped together in the [energy](#) cluster, and they are all close to each other in a 2-dimensional PC space. Also the words [drilling](#) and [exploration](#) are grouped together and this cluster is located very close to the [energy](#) cluster. On the other hand, words such as [loan](#), [loans](#), [mortgage](#), [credit](#), and [bank](#) are clustered together (labeled as [loan](#) cluster) and they are in the other side of the space, indicating the dissimilarity to the [energy](#) and [drilling](#) clusters. Also, words [internet](#), [services](#), [software](#), [solutions](#), and [wireless](#) constitute another cluster labeled as [software](#) which is located as far as possible to clusters [energy](#) and [loan](#).

Most of these main risky words have positive average coefficients in the LASSO cross-sectional regressions, implying that higher usage of these words corresponds to higher returns. Some of these words, however, display a negative average coefficient. For instance, the words [loan](#), [loans](#) and [mortgage](#) are on average associated with lower future returns. Similarly, the more usage of words [China](#) and [business](#) lead, on average, to lower returns. I have marked

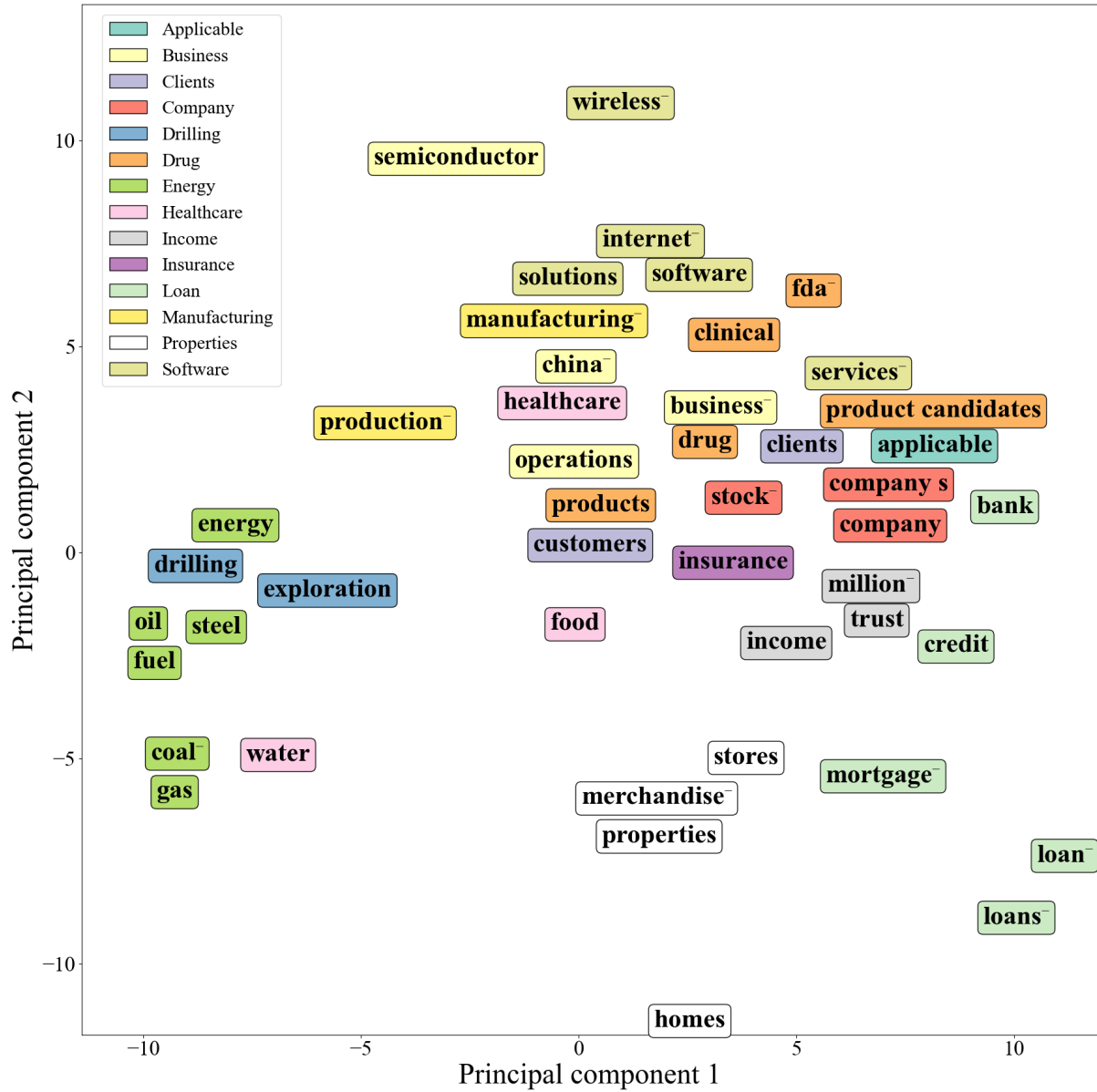


Figure 2. Projecting the main risky words into a 2-dimensional PC space

The figure illustrates the clustering of main risky words based on their vector representations. Each word is located on a 2-dimensional graph based on the first two principal components (PC) derived from their high-dimensional embeddings. Words that are located closer to each other in a 2-dimensional PC space share similar semantic meaning to each other. Using agglomerative clustering with a defined distance threshold, words with high similarity are grouped into 14 clusters shown with the same color. The cluster labels include applicable, business, clients, company, drug, drilling, energy, healthcare, income, insurance, loan, manufacturing, properties, and software. For example, terms like **coal**, **fuel**, **oil** and **energy** are part of the "energy" cluster, while **loan**, **credit**, **mortgage** form the "loan" cluster, positioned far from "energy," indicating significant dissimilarity. Words marked with negative coefficients, such as **loan** and **business**, imply associations with lower returns.

those words with a negative average coefficient in Figure 2.

C. Clustering all the Risky Words

After clustering the 46 main risky words, I then use these as the cores of each cluster, to cluster the rest of the "less important" risky words. There are 1152 risky words that less than

60 times have non-zero coefficients in the LASSO cross-sectional regressions. I assign each of these words to one of the main clusters based on the similarity of their vectors to each member of the main risky words in each cluster. For instance, the words **medical** and **hospital** are most close to the **healthcare** and hence end up in this cluster. Similarly, the words **digital** and **technology** belong to the **software** cluster. Figure 3 shows the words for each cluster that have been repeated at least 12 times in the LASSO cross-sectional regressions. Each word size in this photo is proportional to the number of times that it gets non-zero coefficients. The blue words have a positive average coefficient while the red words' average coefficient is negative. Appendix C lists all the words and their repetitions in each cluster.

To validate how much risky words are similar (dissimilar) within (across) each cluster, I report the average cosine similarity between clusters in Table A1 in Appendix. The diagonal of the similarity matrix shows the high values, meaning that words within each cluster have a high semantic similarity. For instance, the cluster words in the **energy** cluster show an average similarity of 24.28%. Words in **insurance** cluster have the highest similarity of 42.40%. On the other hand, words across clusters have fairly low similarities. For instance, the similarity of cluster **drug** to cluster **loan** is -0.65%, implying no relationship between these words. Also the relationship between the cluster **drug** and **healthcare** is as low as 1.97%.

D. Predictive Power of Risky Words

I study the time-varying predictive power of risky words in several ways. First, I study the average coefficients of words in each cluster over time. Then I study the correlation of the predicted values from each cluster with the overall predicted returns. Lastly, I decompose the variance of predicted returns to the variance of each cluster.

To begin with, I define the coefficient of each cluster at each month as the average coefficients of individual words in that cluster. Then, I plot the 6-month moving average of each cluster coefficient in Figure 4. Figure 4 shows that the coefficients are highly time-varying and they move in cycles. However, the time-series average of coefficients for all clusters are positive, meaning that investors are rewarded by a positive premium over the period of 2002-2023. Cluster **drug** has the highest average coefficients. With respect to average absolute values, clusters **drilling**, **drug** and **energy** have the highest coefficients, suggesting that these clusters are strongest predictors of returns.

To confirm the cyclical behavior of coefficients in each cluster, I employ two tests. The first is the Hodrick-Prescott (HP) filter, which is primarily used to separate a time series into its trend and cyclical components. The output indicates that for each cluster, the standard deviation of the cyclical component is greater than the standard deviation of the trend component. This suggests that the variability in the coefficients is primarily driven by short-term cyclical movements rather than long-term trends.

To understand the dominant frequency in the cycles, I utilize the Fast Fourier Transform (FFT) on the coefficients. I find that the clusters **applicable**, **business**, **company**, **drug**, **energy**, **income**, and **properties** have cycles of approximately half a year, around 6–7 months. On the other hand, the clusters **clients**, **drilling**, **healthcare**, **insurance**, and **loan** show cycles of approximately a quarter, around 3–4 months. Lastly, the clusters **manufacturing** and **software**

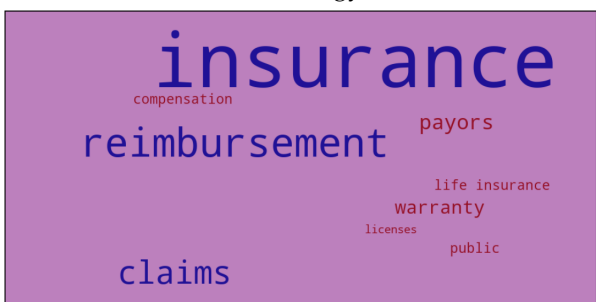
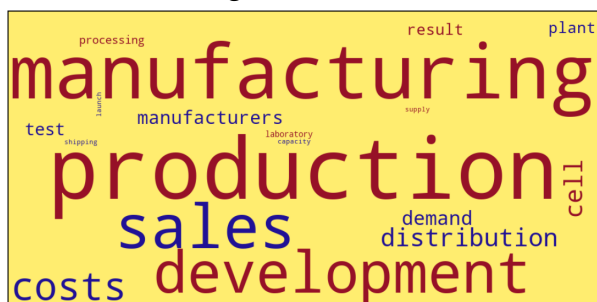
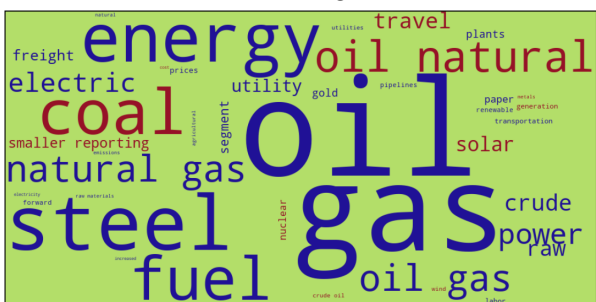
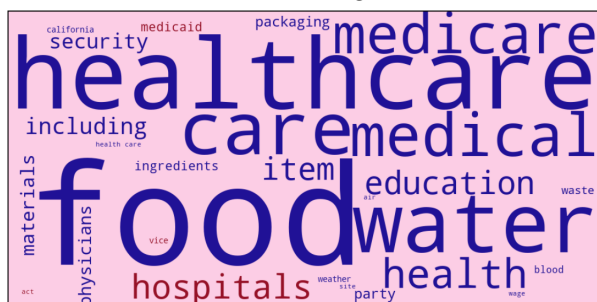
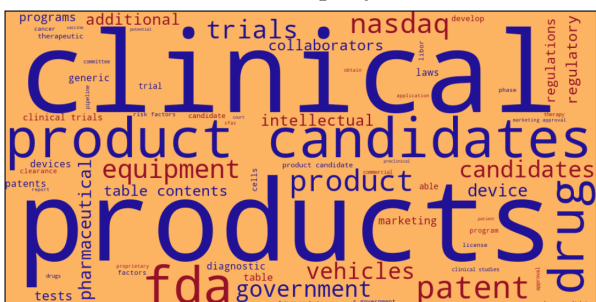
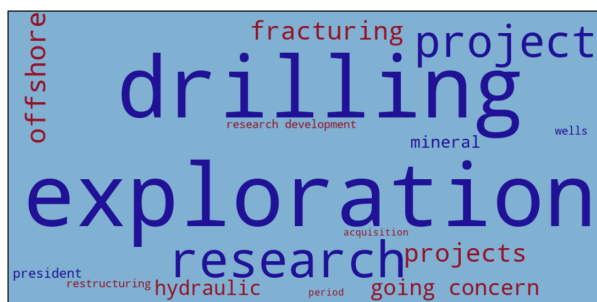
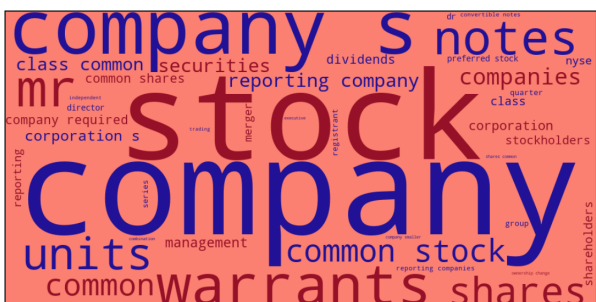
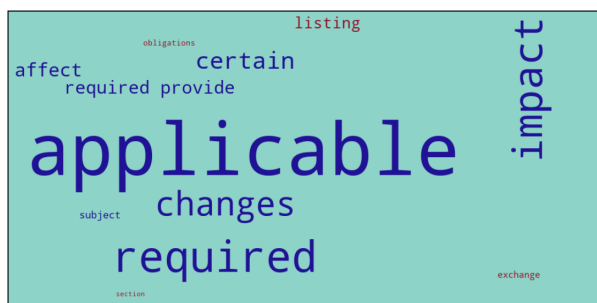


Figure 3. Word clouds of different clusters (continued on next page)

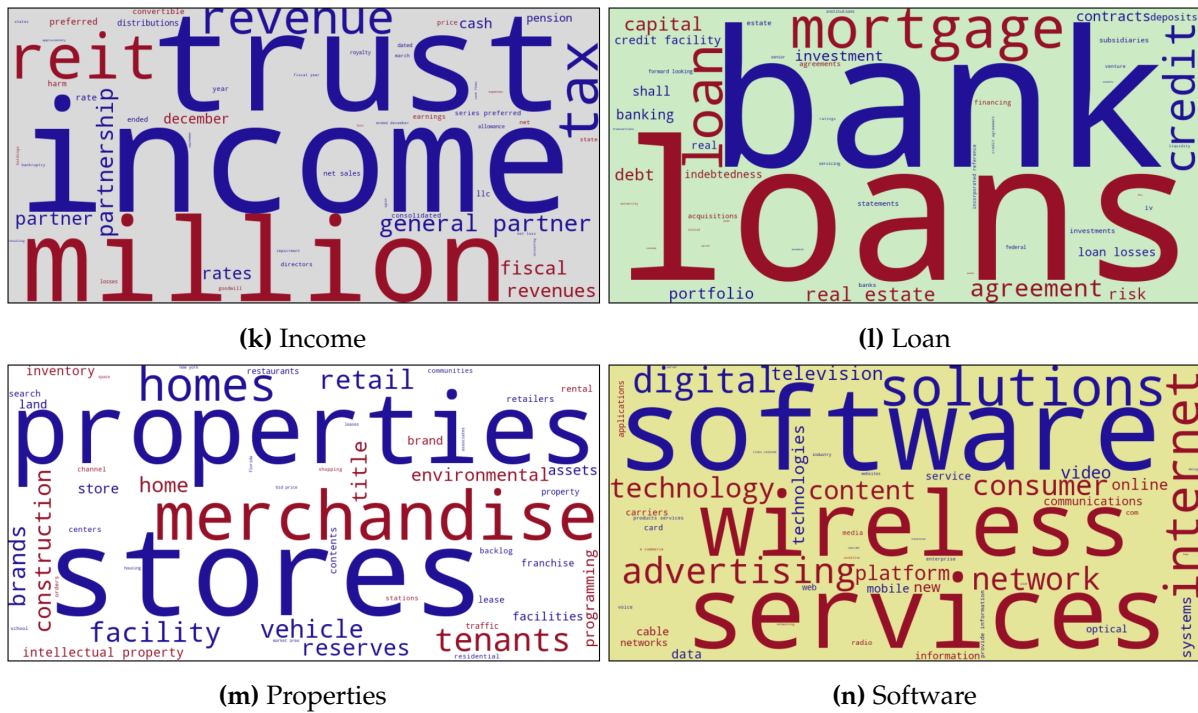


Figure 3. Word clouds of different clusters (continued)

This figure presents a word cloud for each cluster of risky words identified through clustering. The clusters are built using 46 primary risky words as cores, and additional 1152 less significant risky words are assigned to these clusters based on vector similarity. Words appearing at least 12 times with non-zero coefficients in the LASSO cross-sectional regressions are shown, with their size proportional to their frequency. Blue words indicate a positive average coefficient, suggesting higher usage is associated with higher returns, while red words indicate a negative average coefficient, implying lower future returns.

exhibit relatively long cycles, nearly a 12-month annual pattern.

Next, I conduct a variance decomposition to check how much of the variations in predicted expected returns emanate from the variation of individual clusters. Figure 5 shows the normalized proportion of the variance for each risk (to show the graph smoother, I use a 6-month moving average while plotting). Figure 5 shows that in a period around 2007-2008, the cluster **energy** drives a large portion of variation in stock returns. During this period, oil prices surged from approximately \$60 per barrel in early 2007 to a peak of \$147.30 in July 2008. Another notable event was the oil price plunge between mid-2014 and early 2016, where prices fell by approximately 70%. During this period the **energy** cluster explains a large proportion of the variance of returns.

Cluster **loan** also explains a large fraction of the variance of returns during the global financial crisis 2007-2008. After this period, the contribution of cluster **loan** promptly diminishes. In early 2000, also a significant amount of the variance is contributed to the **software** cluster. This can be due to the rise and advancement of the internet, software, and mobile technology during that period.

As the predictions for each stock return come from different sources of risks, the expected returns can be written as the sum of predicted values from each risk:

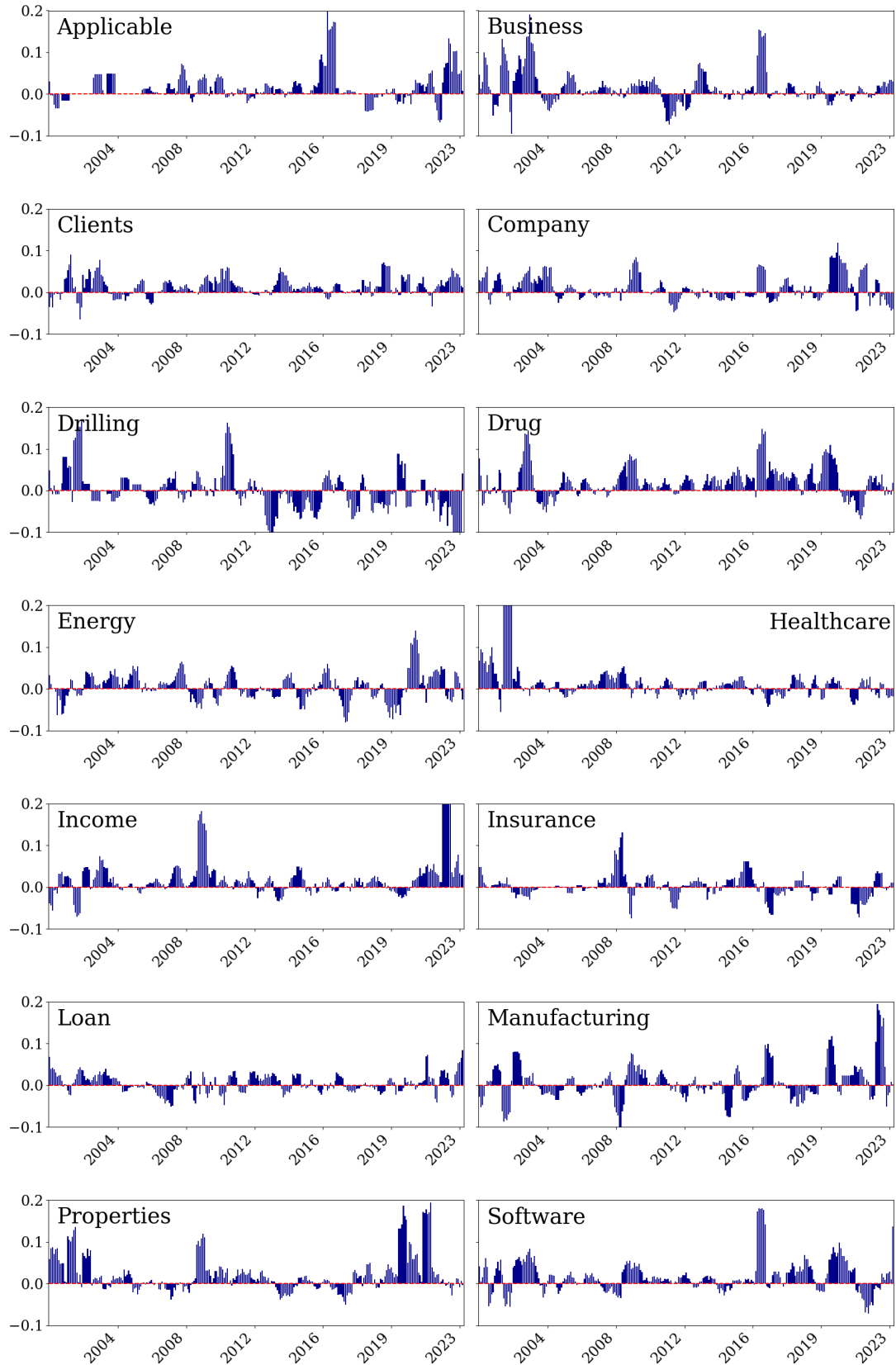


Figure 4. Time-series coefficients of each cluster

This figure shows the time-series 6-month moving average of coefficients of each cluster. The coefficient of each cluster is the average coefficients of words on that cluster from LASSO cross-sectional regressions.

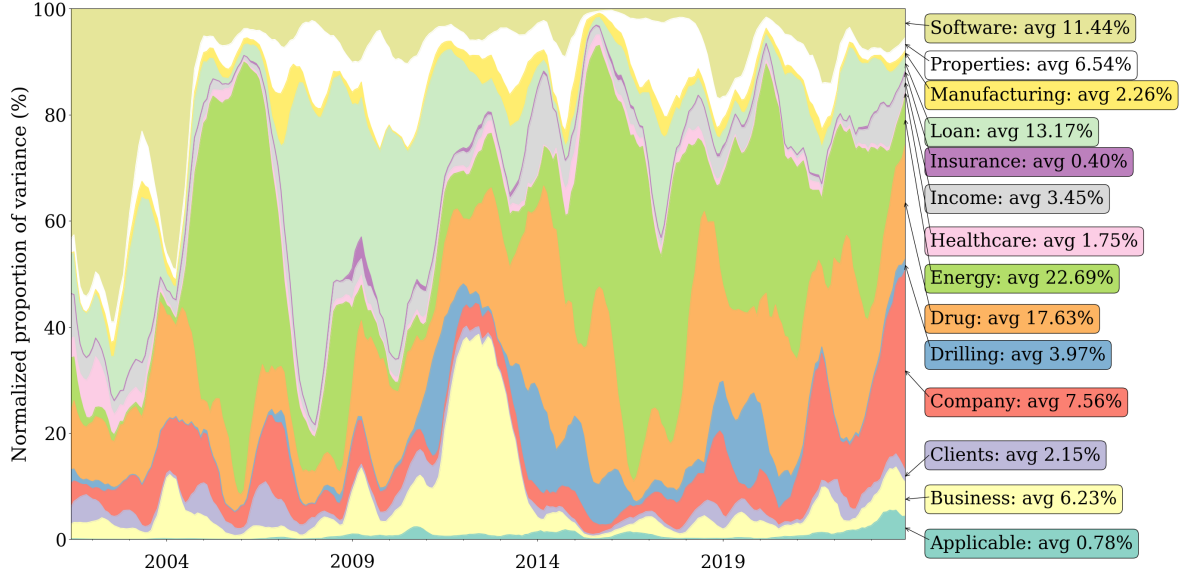


Figure 5. Decomposing variance of predicted returns

In this figure, I present the decomposition of variance in predicted returns across different risky clusters over time, using a 6-month moving average for smoother visualization. The figure shows how much of the variation in predicted returns comes from the variation in each cluster, relative to the variance of all other risky clusters. The **energy** cluster contributes significantly to the variance during the 2007-2008 period, during the sharp oil price drop between 2014 and 2016. The **loan** cluster explains a large proportion of the variance during the 2007-2008 global financial crisis. In the early 2000s, I note that the **software** cluster accounts for a substantial portion of the variance, likely due to the rapid rise of the internet, software, and mobile technology during that period.

$$\tilde{r}_{i,t} = \underbrace{\sum_{\text{word}_j \in \text{energy}} b_{j,t} \text{word}_{j,i,t}}_{\text{cluster energy}} + \underbrace{\sum_{\text{word}_j \in \text{drug}} b_{j,t} \text{word}_{j,i,t} + \dots}_{\text{cluster drug}} + \underbrace{\sum_{\text{word}_j \in \text{company}} b_{j,t} \text{word}_{j,i,t}}_{\text{cluster company}}. \quad (6)$$

Using this equation, I isolate each risk and make predictions for future returns based on each cluster separately. Table 3 shows the performance of long-short portfolios that focus on only one risk at a time. Note that this analysis might be considered an in-sample analysis as I form clusters derived from vectors based on all risk sections. Comparing results from Table 3 to Table 1 where I use all risks simultaneously, confirms that although each risk in isolation has some predictive power for returns, the long-short portfolio's performance is at least 50% higher when the risky clusters are used jointly. The clusters with the highest predictive power include **drilling** and **energy** where they generate more than 1% CAPM alpha per month. Next important clusters include **loan**, **properties**, **manufacturing**, and **company**. Cluster **applicable** seems to have no predictive power when considered in isolation.

Another observation is that the predicted values from different clusters exhibit a surprisingly low correlation. Table 4 shows the average correlation matrix between the predicted values from each cluster, and between the overall predicted returns (column under risky words) and actual realized returns (column under realized returns). The overall predicted values from all risky words show a correlation of 2.79% with the realized returns. Cluster **energy** shows the

	mean	α_{CAPM}	α_{FF4}	α_{FF8}
drilling	1.04*** [2.81]	1.35*** [3.79]	1.23*** [3.68]	1.05*** [3.19]
energy	0.91*** [2.87]	1.13*** [3.69]	0.96*** [3.46]	0.99*** [3.70]
loan	0.78*** [2.97]	0.83*** [3.16]	0.70*** [3.04]	0.64*** [2.74]
properties	0.55** [2.45]	0.58** [2.55]	0.51** [2.47]	0.40* [1.95]
clients	0.54*** [2.99]	0.55*** [2.98]	0.48*** [2.77]	0.36** [2.04]
manufacturing	0.53* [1.75]	0.78*** [2.69]	0.67** [2.43]	0.57** [2.11]
company	0.52*** [3.03]	0.66*** [3.92]	0.59*** [3.74]	0.53*** [3.34]
income	0.39* [1.86]	0.46** [2.20]	0.41** [2.14]	0.35* [1.83]
drug	0.37 [1.33]	0.44 [1.56]	0.35 [1.27]	0.35 [1.26]
insurance	0.30 [1.02]	0.64** [2.43]	0.58** [2.28]	0.42 [1.65]
software	0.29 [1.60]	0.42** [2.42]	0.33** [2.08]	0.22 [1.44]
business	0.28 [1.17]	0.51** [2.23]	0.38* [1.85]	0.19 [0.97]
healthcare	0.25* [1.78]	0.25* [1.76]	0.23 [1.64]	0.16 [1.09]
applicable	0.07 [0.38]	0.13 [0.65]	0.10 [0.56]	-0.02 [-0.13]

Table 3- Predicting returns using each risky cluster

The table displays the performance of long-short portfolios constructed by isolating each cluster risk factor. That is, for predicting future returns, I only use the words related to only one cluster. Then I sort stocks into equally-weighted decile portfolios and construct a long-short strategy from 2005 to 2023. The results show how much each cluster has predictive power for future returns. The analysis can be considered as an in-sample because I form clusters based on the vectors that are constructed by the text of all risk sections. The first column shows the monthly average excess returns. The second column is the CAPM monthly alpha, and the third and fourth columns are the Fama-French-Carhart 4 factor alpha and Fama-French 5 augmented by momentum, short and long-term reversals factors (FF8). The *t*-stats are shown in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

	Realized Returns	Risky Words	Applicable	Business	Clients	Company	Drilling	Drug	Energy	Healthcare	Income	Insurance	Loan	Manufacturing	Properties
Risky Words	2.79														
Applicable	0.27	10.13													
Business	0.40	30.11	1.29												
Clients	0.82	20.35	4.63	1.79											
Company	0.97	25.05	1.23	8.88	-0.21										
Drilling	1.28	27.45	0.77	5.30	1.92	2.32									
Drug	0.37	41.32	1.77	1.96	2.47	2.06	3.15								
Energy	1.79	44.39	0.47	2.80	1.86	-0.92	17.25	3.49							
Healthcare	0.45	14.03	2.58	-1.11	4.86	0.38	-0.31	2.80	-1.35						
Income	0.67	26.97	7.05	6.17	5.16	2.08	1.79	2.88	1.34	-0.88					
Insurance	0.51	9.66	2.44	3.63	-1.58	1.59	2.01	0.38	0.41	8.28	2.07				
Loan	1.28	36.09	3.96	9.05	4.86	-1.45	2.81	2.21	1.47	1.18	9.74	2.14			
Manufacturing	1.14	28.69	2.67	6.77	3.23	1.54	13.22	8.42	11.40	1.25	6.05	4.18	4.04		
Properties	0.64	30.25	-0.44	3.92	-0.86	1.41	3.22	1.75	3.07	2.92	9.70	2.42	7.97	6.21	
Software	0.61	34.22	2.39	6.63	8.73	4.21	1.28	1.11	1.76	1.34	6.04	3.79	4.50	6.70	5.99

Table 4- Correlation between predicted values from each cluster

The table presents the correlation matrix between the predicted values from each cluster and between overall predicted returns (column under "risky words") and actual realized returns (column under "realized returns").

highest correlation with both realized returns and risky words compared to other clusters. The predicted values from clusters **energy** and **manufacturing** show a correlation of 11.40%, while the correlation between **energy** and **drilling** is 17.25%. The clusters **energy** and **healthcare** show a correlation of -1.35%. The average correlation between all clusters is as low as 3.42%, implying that the predicted values from each cluster are fairly independent.

E. Excerpts of the Risky Words

In this part, I present a summary of some of the clusters to shed light on what firms on average write about when mentioning essential risky words. I use the word embedding model to summarize the excerpts. For each risky word, I first find all sentences in risk sections in months that this word has a non-zero coefficient. Then I use the word embedding model to extract the core idea of these sentences.

My approach begins with converting the text of each document into vector representations. Given the document contains all excerpts of the risky word, I calculate the vector for the entire document as the average vector for all individual words. This leads to a single vector that represents the overall content and theme.

The average vector acts like a centroid in the vector space, indicating the general direction

and focus of the document's meaning. To extract a representative sentence that best reflects the core ideas of the excerpts, I calculate the cosine similarity between the average vector and the vectors of each sentence in the document. The sentence with the highest cosine similarity score is the one that most closely matches the average vector. Selecting this sentence as the summary ensures that I capture the most relevant and central idea in the document, resulting in a concise and meaningful representation of the original text.

Below I present the most similar sentences to the average for three words *oil*, *clinical*, and *stock*.

- **Risky word:** *oil*
 - **Cluster:** Energy.
 - **Sentence** (similarity 0.91): "In addition, the market prices and demand for *oil* and natural gas are impacted by governmental regulations and the level of *oil* and natural gas production in the United States and non-OPEC+ countries, as well as the *oil* and gas industry's view of future *oil* and gas prices, which generally determine the level of capital spending for the exploration, development and production of crude *oil* and natural gas reserves."
- **Risky word:** *clinical*
 - **Cluster:** Drug.
 - **Sentence** (similarity 0.93): "We also may experience numerous unforeseen events during, or as a result of, any future *Clinical* Studies or *Clinical* Trials that we may conduct that could delay or prevent our ability to receive marketing approval or commercialize our product candidates, including, but not limited to: unforeseen events or events over which we have little to no control, such as the COVID-19 pandemic, can cause execution delays for our *Clinical* Studies or *Clinical* Trials; issues related to the quality, completeness and interpretability of our data that could result in significant delays or additional costs and impact development plans for our product candidates; we may be unable to generate sufficient pre*clinical*, toxicology or other in vivo or in vitro data to support the initiation of *Clinical* Trials for therapeutic indications for any drug product candidates or the marketing of our products as non-drug products; results from one *Clinical* Study or *Clinical* Trial, particularly observation of a serious adverse event, may impact the other *Clinical* Study or *Clinical* Studies or *Clinical* Trial or *Clinical* Trials of the same product candidate; the FDA may not allow us to use data from our *Clinical* Studies to support a late-phase IND *Clinical* Trial or an IND *Clinical* Trial of any phase for a product candidate we decide to develop as a drug product candidate instead of a non-drug product candidate; the FDA or other regulatory authorities may disagree with the design, implementation or results of our *Clinical* Studies or *Clinical* Trials, which may delay or prevent us from pursuing certain regulatory pathways for product developments, or require us to submit additional data such as long-term toxicology studies or impose other requirements before permitting us to initiate or complete a *Clinical* Trial of any phase."

- **Risky word:** *stock*
 - **Cluster:** Company.
 - **Sentence** (similarity 0.93): “If we issue any shares of our Common *stock* or securities convertible into our Common *stock* in the future, including the issuance of shares of Common *stock* upon conversion of any convertible notes or the issuance of shares of Common *stock* upon exercise or settlement of any outstanding *stock* options, restricted *stock* units or performance *stock* units granted under the PHH Corporation Amended and Restated 2005 Equity and Incentive Plan, such issuances will dilute the voting power and ownership percentage of our *stock*holders and could substantially decrease the trading price of our Common *stock*.”

VI. Additional Results

In this part, I discuss additional results including controlling for industries, firm characteristics, sentiments, and text features. I further test asset pricing models with risky words. I also address the concerns of data mining and alternative approaches.

A. *Risky words and industries, characteristics, sentiments, and text features*

A.1. Risky words and industries

Because some of the clusters that I find resemble the standard industry classes, an important concern is to check if risky words are just proxies for industries. For instance, if only firms in the energy industry mention energy-related words, then risky words would be fully captured by firm industries. However, I show that this is not the case: industries do not capture the risky words strategy. I also show that firms in different industries are exposed to all of the risky clusters, and words in different clusters are mentioned by all firms.

To begin with, I assign each firm to one of the 10 industry classes of Fama and French (later I expand to 48 classes for a more refined analysis). Then I measure how much each firm in each industry mentions words in each cluster. Then I aggregate the results over cluster and time (between 2005-2023). Figure 6 part (a) shows the shares of each risky cluster in each industry—how much firms in each industry allocate their risk sections to each risky cluster over time. The x-axis shows different industry classes such as “Durbl” (Durables), “Enrgy” (Energy), “HiTec” (High Technology), “Hlth” (Healthcare), “Manuf” (Manufacturing), “NoDur” (Non-Durables), “Other,” “Shops,” “Telcm” (Telecom), and “Utils” (Utilities) and the y-axis represents the percentage share of mentions for each risky cluster within the total risk section of the industry. Each color block in the stacked bars represents a different risky cluster (e.g., *applicable*, *business*, *clients*, etc.)

Consider for instance the energy industry. A significant share of risk is allocated to clusters like *energy* (18.7%) and *drilling* (18.3%). However, the rest of 63% of risk sections is allocated to other risky clusters. For the Healthcare (Hlth) industry, the *drug* cluster (16.8%) and *healthcare* (10.2%) are more dominant, but there are equally much of discussions on *manufacturing* cluster (9.11%), and *business* (9.75%). For the manufacturing industry, only 10.9% of the used risky

words belong to **manufacturing** cluster. The telecom industry chooses only 13.5% of its words from **software** industry. Most industries show diversity, with a relatively even distribution across multiple clusters. This figure shows that the risky clusters are fairly irrelevant to standard industry classes.

Part (b) of Figure 6 offers a reverse view. I show how words in each risky cluster are distributed over industries. The x-axis indicates the percentage of mentions that each industry contributes to a given risk cluster. The y-axis lists different risky clusters.

Another way to check for industries is to sort stocks based on the predicted values from risky words within each industry. Table A2 in the Appendix shows the results where I sort stocks into three portfolios based on the predicted value of risky words, and create a long-short 3-minus-1 portfolio within each industry. The left side of Table A2 shows the mean excess returns of each portfolio, while the right side shows the alpha with respect to the Fama-French 8-factor model. The table suggests that almost in all industries risky words predict the cross-sectional variation in average returns.

Each risky cluster has a fairly similar distribution of industries contributing to it, although some industries might be more dominant in a relevant cluster. For example: in the **drilling** cluster the “Energy” industry (27.7%) dominates, highlighting the relevance of energy-related risk in drilling. Or, in the **drug** cluster the “Healthcare” industry (24.7%) is prominent, which is reasonable given the drug-related risks in healthcare sectors. Some clusters, like **manufacturing** and **business**, show a more even distribution across multiple industries, suggesting that these risks are broadly relevant.

Next, I rebuild the risky-words portfolios while adjusting for industries. That is, I subtract the average returns of each industry from each asset. Table 5 shows the industry-adjusted performance of each portfolio. The left side of the table shows the case where all stocks are used for creating portfolios. A long-short portfolio generates an average return of 0.73% ($t = 2.96$) per month. The CAPM alpha is 0.84% ($t = 3.45$). This portfolio generates economically large and statistically significant alphas with respect to the Fama-French four and eight factor models. The right side of Table 5 shows the counterpart results for all but tiny stocks, confirming that the results do not stem merely from microcaps.

A.2. Double sort portfolios

Another important consideration is to rule out the possibility that the risky words strategy simply proxies for some firm characteristics, sentiments, or other text features. I check this possibility by both double sorting stocks into portfolios and also standard Fama-MacBeth regressions.

Because I obtain the risky word coefficients from the past five years and use those coefficients to predict next month’s returns, an obvious concern is if past returns can explain the risky words strategy performance. To check this, I independently create double-sort portfolios based on the risky words prediction and momentum (also short and long-term reversals), and other characteristics. More specifically, similar to Florackis et al. (2023), I conduct double sorting starting from January 2005, where stocks are sorted at the end of each month in ascending order based on their predicted value from risky words and categorized into three groups: 1,

	All Stocks				All But Tiny Stocks			
	mean	α_{CAPM}	α_{FF4}	α_{FF8}	mean	α_{CAPM}	α_{FF4}	α_{FF8}
1	-0.51*** [-2.82]	-0.61*** [-3.45]	-0.45*** [-3.11]	-0.36** [-2.51]	-0.32* [-1.88]	-0.48*** [-3.03]	-0.31*** [-2.68]	-0.23** [-2.07]
2	-0.32*** [-3.51]	-0.36*** [-3.99]	-0.29*** [-3.74]	-0.27*** [-3.39]	-0.09 [-1.06]	-0.15* [-1.87]	-0.09 [-1.30]	-0.05 [-0.70]
3	-0.15** [-2.20]	-0.18** [-2.59]	-0.15** [-2.24]	-0.14** [-2.01]	-0.08 [-1.27]	-0.09 [-1.45]	-0.08 [-1.35]	-0.11* [-1.70]
4	0.01 [0.23]	0.01 [0.23]	0.01 [0.21]	0.01 [0.20]	-0.02 [-0.39]	-0.01 [-0.24]	-0.01 [-0.20]	0.01 [0.21]
5	0.02 [0.33]	0.02 [0.36]	0.01 [0.10]	-0.00 [-0.01]	0.12* [1.91]	0.13** [2.20]	0.11* [1.88]	0.11* [1.74]
6	0.10 [1.56]	0.12** [1.99]	0.10* [1.70]	0.07 [1.17]	0.01 [0.19]	0.03 [0.55]	0.00 [0.07]	-0.03 [-0.56]
7	0.24*** [3.70]	0.26*** [3.95]	0.22*** [3.63]	0.20*** [3.08]	0.12* [1.95]	0.17*** [2.90]	0.14** [2.54]	0.12** [2.07]
8	0.18** [2.39]	0.23*** [3.15]	0.18*** [2.81]	0.13** [2.05]	0.16** [2.55]	0.21*** [3.44]	0.17*** [3.10]	0.13** [2.40]
9	0.22*** [2.70]	0.28*** [3.54]	0.22*** [3.18]	0.20*** [2.91]	0.01 [0.16]	0.07 [0.94]	0.02 [0.38]	-0.00 [-0.06]
10	0.22* [1.90]	0.23** [1.98]	0.17* [1.66]	0.17* [1.76]	0.08 [0.69]	0.11 [1.02]	0.03 [0.35]	0.04 [0.42]
LS	0.73*** [2.96]	0.84*** [3.45]	0.61*** [3.30]	0.53*** [2.94]	0.40 [1.63]	0.60** [2.56]	0.34** [2.08]	0.27* [1.70]

Table 5- Industry adjusted returns of risky words portfolios

This table shows the monthly out-of-sample performance of the portfolios based on the risky words strategy where the portfolio returns are industry-adjusted. I identify the industries based on the SIC code from Fama French 10 industry classes. When creating portfolios, I subtract the realized returns of each asset from its average industry at each month. The four left columns show the results when using all assets, while the four right columns show the results when excluding the market caps. All portfolios are equally weighted. Column mean shows the average excess returns of each portfolio. Other columns show the alpha with respect to CAPM, the Fama-French-Carhart 4 factor model, and Fama-French 5 augmented by momentum, short and long-term reversals factors (FF8). The last row shows a long-short portfolio (LS). The time period is between 2005-2023. The t -statistics are shown in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

2 and 3. Additionally, I perform independent sorting of stocks in ascending order according to various firm-level and 10-K document characteristics. These stocks are divided into three portfolios (low, middle, and high) using the corresponding percentiles for characteristics such as momentum, short-term reversal, long-term reversal, size, book-to-market, beta, idiosyncratic volatility, length of the risk sections, readability of the risk section, its sentiment, and changes in the risk sections. The combination of these independent sorts results in multiple double-sorted portfolios, whose performance is then monitored over the subsequent month. The outcomes are detailed in Table 6, where I report the excess returns of portfolios with high predictions from the risky words compared to those with risky words predictions across different subsamples.

Notably, higher returns for stocks with high prediction based on risky words are observed consistently across all subsamples and remain statistically significant in most cases. This confirms that my results are not limited to a specific subset of stocks and addresses concerns that risky words might be correlated with other known risk factors.

A.3. Fama-MacBeth regressions

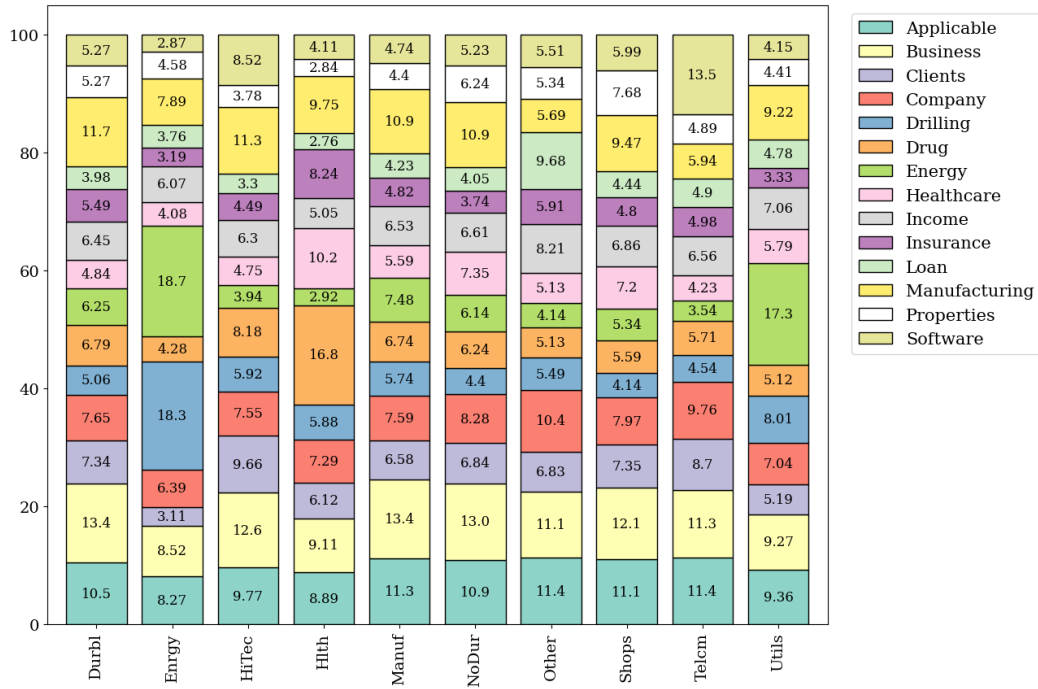
Next, I perform cross-sectional [Fama and MacBeth \(1973\)](#) regressions to control for multiple effects of many predictors at the same time. I use the predicted value based on the risky words as "risk score" and include other well-known predictors in the regression. At each month, I regress the stock returns on the lagged risk score and additional characteristics. I include momentum, short-term and long-term reversal, size, book-to-market, beta, illiquidity, coskewness, idiosyncratic volatility, firm age, maximum daily return, R&D, asset growth, length of the risk section, readability, number of positive and negative words based on the machine learning dictionary of [García et al. \(2023\)](#), and changes in the risk sections as suggested by [Cohen et al. \(2020\)](#). I also control for the 48-industry classification of Fama and French. Table 7 shows the results of Fama-MacBeth regressions along with t -values obtained based on Newey-West standard errors.

In the first column of Table 7, I regress one month ahead returns on the most recent characteristics. The coefficient for the risk score is 2.85 with $t = 7.02$. This is the highest t -value among all the regressors. In the second column, I regress two-month ahead on the same variables. The t -value for the coefficient of the risk score is 5.21. In the third column onwards, I use the three, four, five, six, nine and twelve-month ahead returns and observe that the risky words predictive power remains strong and significant. This suggests that the risky words' predictive power is long-term.

B. Data mining, overfitting or actual risk?

Risk disclosures are quite lengthy and noisy which raises two important concerns: data mining and overfitting.

Data mining occurs when patterns are found simply by chance due to a large number of variables rather than because of any underlying genuine relationship. The large dimensionality of text data indeed provides ample opportunity to find spurious correlations. Overfitting is another core risk: with too many potential predictors, there is a danger that the model might fit to the noise in the data rather than capturing the real underlying patterns, meaning it would perform well on the training data but poorly when exposed to new data.



(a) Shares of each risky cluster in each industry



(b) Shares of each industry in each risky cluster

Figure 6. Distribution of risky clusters and industries

This figure presents an analysis of how risky clusters are associated with various industries, depicted in two complementary parts. Part (a) shows the shares of each risky cluster within each industry, demonstrating how firms allocate mentions of different risk-related topics over time, from 2005 to 2023. Each vertical bar represents an industry categorized under the Fama and French 10-industry classification, with colored segments illustrating the proportion of mentions dedicated to each risk cluster. In contrast, Part (b) reverses this perspective by showing the distribution of industries contributing to mentions within each risky cluster. Each horizontal bar represents a cluster, and the colored segments indicate the share of each industry within that cluster. This part reveals which industries are most relevant to or contribute significantly to each type of risk.

	mean				α_{FF8}			
	1	2	3	LS	1	2	3	LS
<i>Momentum</i>								
low	0.09	0.80	1.29**	1.20***	-0.37*	0.24	0.71***	1.08***
	[0.16]	[1.57]	[2.48]	[4.19]	[-1.70]	[1.57]	[3.30]	[4.25]
high	0.61	0.95***	1.12***	0.50***	0.11	0.24***	0.38***	0.27
	[1.59]	[2.65]	[2.92]	[2.63]	[0.90]	[2.76]	[3.28]	[1.65]
H-L	0.52	0.15	-0.18		0.48**	-0.00	-0.33*	
	[1.59]	[0.53]	[-0.63]		[2.35]	[-0.02]	[-1.79]	
<i>Short term reversal</i>								
low	0.48	1.03**	1.35***	0.87***	-0.01	0.32**	0.64***	0.66***
	[0.86]	[2.17]	[2.80]	[3.17]	[-0.08]	[2.28]	[3.62]	[2.88]
high	-0.07	0.61	1.06***	1.13***	-0.49***	0.02	0.41***	0.91***
	[-0.17]	[1.62]	[2.71]	[5.07]	[-3.06]	[0.21]	[3.00]	[4.78]
H-L	-0.55**	-0.41*	-0.29		-0.48***	-0.30**	-0.23	
	[-2.18]	[-1.92]	[-1.38]		[-2.71]	[-2.11]	[-1.56]	
<i>Long term reversal</i>								
low	0.30	1.08**	1.46***	1.16***	-0.10	0.48***	0.80***	0.90***
	[0.56]	[2.36]	[3.07]	[4.04]	[-0.44]	[3.12]	[4.02]	[3.77]
high	0.41	0.75**	1.01***	0.60***	-0.18	0.05	0.29***	0.46***
	[0.93]	[2.09]	[2.80]	[2.80]	[-1.45]	[0.59]	[3.16]	[2.89]
H-L	0.11	-0.33	-0.45*		-0.08	-0.43**	-0.52***	
	[0.43]	[-1.49]	[-1.85]		[-0.35]	[-2.43]	[-2.68]	
<i>Size</i>								
low	-0.02	0.97**	1.54***	1.56***	-0.34	0.52**	1.13***	1.47***
	[-0.04]	[2.13]	[3.24]	[5.63]	[-1.32]	[2.22]	[4.21]	[5.94]
high	0.67*	0.87**	0.95***	0.28	0.09	0.10	0.15	0.06
	[1.67]	[2.44]	[2.74]	[1.28]	[0.74]	[0.85]	[1.62]	[0.38]
H-L	0.69**	-0.10	-0.59*		0.43	-0.43	-0.98***	
	[2.45]	[-0.33]	[-1.90]		[1.61]	[-1.55]	[-3.60]	

Continued on next page

	mean				α_{FF8}			
	1	2	3	LS	1	2	3	LS
<i>BM</i>								
low	0.41 [0.86]	0.73* [1.86]	0.93** [2.30]	0.52** [2.14]	-0.10 [-0.70]	0.09 [0.96]	0.21* [1.75]	0.32 [1.60]
high	0.15 [0.32]	1.06*** [2.81]	1.46*** [3.43]	1.31*** [4.20]	-0.25 [-1.24]	0.48*** [3.53]	0.83*** [4.31]	1.08*** [4.06]
H-L	-0.26 [-1.24]	0.33** [1.97]	0.53** [2.46]		-0.15 [-0.80]	0.40*** [2.92]	0.61*** [3.34]	
<i>Beta</i>								
low	0.23 [0.71]	0.73** [2.51]	1.10*** [3.80]	0.87*** [4.38]	-0.10 [-0.67]	0.17 [1.43]	0.52*** [4.89]	0.62*** [3.53]
high	0.28 [0.45]	0.97* [1.85]	1.32** [2.46]	1.04*** [3.71]	-0.31 [-1.58]	0.27* [1.85]	0.53*** [2.81]	0.84*** [3.71]
H-L	0.05 [0.12]	0.24 [0.75]	0.22 [0.70]		-0.21 [-0.98]	0.10 [0.54]	0.01 [0.07]	
<i>IdioVol</i>								
low	0.76** [2.27]	0.89*** [2.76]	1.07*** [3.49]	0.31** [2.07]	0.24** [2.20]	0.19* [1.73]	0.36*** [4.62]	0.12 [0.92]
high	-0.19 [-0.32]	0.82 [1.56]	1.13** [2.12]	1.32*** [4.40]	-0.56** [-2.31]	0.31 [1.47]	0.52** [2.11]	1.09*** [4.38]
H-L	-0.95** [-2.60]	-0.08 [-0.23]	0.06 [0.18]		-0.81*** [-2.92]	0.12 [0.48]	0.16 [0.61]	
<i>word count</i>								
low	0.57 [1.38]	0.86** [2.35]	1.25*** [3.33]	0.68*** [3.03]	0.06 [0.41]	0.26*** [3.03]	0.60*** [3.99]	0.54*** [2.99]
high	0.22 [0.40]	0.80* [1.87]	1.23*** [2.76]	1.01*** [3.27]	-0.28 [-1.43]	0.15 [1.01]	0.59*** [3.47]	0.86*** [3.53]
H-L	-0.35 [-1.63]	-0.06 [-0.34]	-0.02 [-0.11]		-0.34* [-1.91]	-0.11 [-0.77]	-0.02 [-0.10]	

Continued on next page

	mean				α_{FF8}			
	1	2	3	LS	1	2	3	LS
<i>Sentiment</i>								
low	0.35	0.95**	1.12***	0.76***	-0.03	0.34***	0.47***	0.50**
	[0.76]	[2.43]	[2.73]	[2.78]	[-0.16]	[2.73]	[2.96]	[2.12]
high	0.28	0.82**	1.18***	0.90***	-0.27	0.16	0.51***	0.78***
	[0.61]	[2.12]	[2.94]	[3.69]	[-1.61]	[1.32]	[4.13]	[3.77]
H-L	-0.07	-0.12	0.06		-0.24	-0.18	0.04	
	[-0.45]	[-0.86]	[0.40]		[-1.47]	[-1.28]	[0.26]	
<i>Changes in the risk section</i>								
low	0.37	0.85**	1.14***	0.76***	-0.07	0.25**	0.45***	0.52**
	[0.81]	[2.19]	[2.88]	[2.95]	[-0.41]	[2.14]	[3.66]	[2.56]
high	0.08	0.69*	1.09***	1.01***	-0.41**	0.07	0.48***	0.89***
	[0.17]	[1.78]	[2.68]	[4.09]	[-2.42]	[0.59]	[3.36]	[4.30]
H-L	-0.30***	-0.16	-0.05		-0.34***	-0.18	0.03	
	[-2.64]	[-1.45]	[-0.47]		[-2.95]	[-1.57]	[0.28]	
<i>Readability</i>								
low	0.44	0.93**	1.13***	0.69**	-0.02	0.26**	0.49***	0.51**
	[0.92]	[2.37]	[2.82]	[2.49]	[-0.11]	[2.06]	[3.67]	[2.14]
high	0.21	0.83**	1.20***	0.99***	-0.33*	0.24**	0.53***	0.86***
	[0.44]	[2.12]	[2.81]	[3.95]	[-1.95]	[2.40]	[2.89]	[4.36]
H-L	-0.23	-0.10	0.08		-0.31*	-0.01	0.04	
	[-1.50]	[-0.76]	[0.47]		[-1.97]	[-0.10]	[0.25]	

Table 6- Double Sort Portfolios

I report average returns and eight-factor alphas from the Fama and French model for double-sorted portfolios based on the risky words strategy and each of the following firm characteristics: momentum, short-term reversal, long-term reversal, size, book-to-market (BM), beta, idiosyncratic volatility, word count, risky words sentiment, changes in the risk section, and readability. From January 2005 to December 2023, I sort stocks at the end of each month in ascending order based on their predicted value from the risky words model and allocate them into three groups. I also independently sort stocks in ascending order by the value of each characteristic listed above, creating three portfolios (here, I only report the low and high portfolios). The portfolios are equally weighted and I rebalance portfolios monthly. The LS is 3-minus-1 and H-L is high-minus low portfolio. t -statistics are reported in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

To mitigate these risks, I apply LASSO regressions. I show that this technique, in combination with cross-validation, is an effective way to prevent overfitting and address the issue of noise. LASSO regression is particularly well-suited for high-dimensional settings where the number of predictors (i.e., words) is large relative to the number of observations (e.g., firm-level returns). The core strength of LASSO is its ability to penalize the coefficients of irrelevant variables, effectively reducing many of them to zero, thus acting as a powerful form of variable selection. By shrinking less useful predictors to zero, LASSO keeps the model parsimonious—only a select subset of words that are meaningful remain, which mitigates the risks of overfitting to noise.

In addition, I use 5-fold cross-validation to ensure that the model is generalizable and not overly tailored to specific data. This approach helps verify that the selected variables (words) provide consistent and robust predictions across different subsets of data, rather than just performing well on a single training set. Together, LASSO and cross-validation create a framework where the model only includes words that have consistently shown predictive power, significantly reducing the likelihood of fitting to noise.

This method leads to a consistent word selection over time, which is strong evidence against overfitting and data mining. For instance, if the model were simply overfitting to random noise, we would expect different sets of words to be selected each time, as noise is by nature inconsistent. The stability of this selection over time demonstrates that the chosen words carry a meaningful signal related to returns, rather than being the result of spurious correlations that might vary from sample to sample. It also indicates that the risky words have genuine predictive value for risk, and that they represent characteristics or concepts present in the disclosures that are materially related to the expected returns.

To further confirm that the words selected are not just coincidentally related to firm-level risk, I perform a random reassignment of risk sections to firms. In other words, I take risk disclosures and assign them to some other firms arbitrarily, and then re-run the model. When the risk disclosures are randomly assigned, the predictive power of the model completely vanishes. No significant words are found. If the words had merely been randomly related to risk, they would likely have shown up even in this incorrectly assigned context. This reassignment functions as a type of placebo test.

I also runs an experiment involving random matrices instead of actual risk disclosures to confirm that LASSO does not just pick predictors from any available data. I generate sparse matrices so that the sparsity is similar to the actual TF-IDF matrices. The average TF-IDF matrices have a sparsity of 85%, meaning that 85% of its elements are zero. The LASSO model does not select random noise as predictors, which reinforces the idea that the original predictors—the specific words selected—have a true, underlying relationship with firm risk. This test demonstrates the selection process is not just an artifact of the model structure itself. It confirms that the original model is detecting actual structure and meaning in the data, rather than being tricked by randomness.

Another observation is that, the number of risky words at each month is independent of the length of the risk sections. The TF-IDF matrix tends to grow larger over time, while the number of risky words picked by the model shows a high correlation with economic uncertainty rather than number of total words. If the risky words were the result of data mining or overfitting

	r_{t+1}	r_{t+2}	r_{t+3}	r_{t+4}	r_{t+5}	r_{t+6}	r_{t+9}	r_{t+12}
Risk score	2.85*** [7.02]	2.22*** [5.21]	2.05*** [4.83]	2.02*** [4.80]	1.57*** [3.55]	1.23*** [2.75]	1.34** [2.50]	1.15** [2.03]
Mom12m	0.02 [0.14]	-0.05 [-0.48]	-0.04 [-0.40]	-0.02 [-0.21]	-0.03 [-0.35]	-0.02 [-0.31]	-0.04 [-0.70]	-0.08 [-1.22]
STreversal	-0.48*** [-4.28]	0.01 [0.20]	0.01 [0.11]	-0.04 [-0.51]	0.08 [1.12]	-0.08 [-1.04]	0.09 [1.30]	0.11* [1.91]
LReversal	0.04 [0.95]	0.03 [0.67]	0.02 [0.40]	0.03 [0.67]	0.04 [0.97]	0.04 [1.12]	0.03 [0.80]	0.01 [0.25]
mvel1	-0.03 [-1.41]	-0.02 [-1.15]	-0.02 [-0.94]	-0.01 [-0.66]	-0.01 [-0.65]	-0.01 [-0.67]	-0.00 [-0.24]	-0.01 [-0.38]
BM	0.10*** [2.64]	0.10*** [2.88]	0.11*** [3.11]	0.09*** [2.66]	0.10*** [2.81]	0.11*** [2.98]	0.07** [2.20]	0.07** [2.06]
Beta	0.03 [0.29]	0.05 [0.53]	0.02 [0.19]	0.02 [0.25]	0.03 [0.36]	-0.01 [-0.09]	-0.06 [-0.88]	-0.04 [-0.57]
Illiquidity	0.17*** [3.19]	0.17*** [3.08]	0.14** [2.39]	0.13** [2.45]	0.12*** [2.80]	0.14*** [3.43]	0.09* [1.73]	0.12** [2.02]
Coskewness	-0.08* [-1.68]	-0.06 [-1.48]	-0.08** [-2.13]	-0.07** [-1.96]	-0.08** [-2.10]	-0.05 [-1.60]	-0.05 [-1.49]	-0.06** [-2.11]
IdioVol3F	-0.46* [-1.69]	-0.07 [-0.31]	-0.23 [-1.25]	-0.17 [-0.74]	-0.10 [-0.41]	-0.08 [-0.28]	0.13 [0.61]	0.02 [0.07]
FirmAge	0.03 [0.60]	0.03 [0.71]	0.05 [1.05]	0.05 [1.15]	0.05 [1.01]	0.04 [0.86]	0.03 [0.71]	0.03 [0.60]
MaxRet	0.15 [0.72]	-0.29* [-1.66]	0.01 [0.07]	-0.04 [-0.26]	-0.13 [-0.74]	-0.15 [-0.67]	-0.29** [-2.06]	-0.20 [-0.98]
RD	0.42*** [3.78]	0.28*** [2.96]	0.21** [2.34]	0.26** [2.53]	0.27*** [2.99]	0.24*** [2.99]	0.24*** [2.65]	0.06 [0.97]
AssetGrowth	-0.15*** [-4.88]	-0.14*** [-4.84]	-0.13*** [-4.61]	-0.13*** [-4.68]	-0.14*** [-4.82]	-0.12*** [-4.47]	-0.11*** [-3.90]	-0.11*** [-3.64]
word count	-0.08 [-1.14]	-0.09 [-1.23]	-0.09 [-1.28]	-0.15** [-2.22]	-0.14** [-2.09]	-0.17** [-2.38]	-0.18*** [-2.88]	-0.20*** [-2.84]
readability score	-0.20* [-1.93]	-0.22** [-2.23]	-0.17 [-1.55]	-0.22** [-2.19]	-0.23** [-2.33]	-0.25** [-2.44]	-0.19** [-2.15]	-0.23** [-2.28]
ML neg	-0.15* [-1.70]	-0.14* [-1.80]	-0.09 [-0.96]	0.04 [0.37]	-0.00 [-0.02]	0.05 [0.52]	0.10 [1.07]	0.19* [1.86]
ML pos	0.23** [2.40]	0.23** [2.55]	0.19** [2.27]	0.15* [1.70]	0.16* [1.68]	0.15 [1.61]	0.12 [1.28]	0.06 [0.63]
changes	-0.34** [-2.36]	-0.31** [-2.15]	-0.31** [-2.28]	-0.36*** [-2.76]	-0.36*** [-2.87]	-0.34*** [-2.73]	-0.32*** [-2.71]	-0.24** [-2.02]
const	1.86*** [3.18]	2.14*** [3.72]	1.86*** [3.26]	2.17*** [3.89]	2.26*** [4.09]	2.47*** [4.39]	2.14*** [4.31]	2.36*** [3.87]
48 Ind FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	769499	765072	760650	756213	751784	747373	734299	721543

Table 7- Cross-sectional Fama-MacBeth regressions

The table presents the results of Fama-MacBeth cross-sectional regressions where one to twelve-month ahead stock returns are regressed on the predicted value from the risky words and various firm characteristics. Each regression includes momentum, short-term and long-term reversal, size, book-to-market, beta, illiquidity, coskewness, idiosyncratic volatility, firm age, maximum daily return, R&D, asset growth, and measures of textual risk (length of risk section, readability, counts of positive (ML pos) and negative (ML neg) words per [García et al. \(2023\)](#)), and changes in risk sections as per [Cohen et al. \(2020\)](#)). The regressions also control for the 48-industry classification of Fama and French, and t -values are computed using Newey-West standard errors. The significance levels are denoted as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

to noise, then one would expect that as the length of the risk sections increases, the model would start selecting more variables. This is because with more text, there are simply more words available to correlate, and in a data mining scenario, the model would likely find a larger number of spurious relationships just by sheer chance. If LASSO was purely overfitting, then probably more variables would have been picked up as the length (and dimensionality) of the text increased, because more noise would provide more opportunities for random correlation. The fact that the number of significant variables remains irrelevant to the length of the document is strong evidence against data mining or overfitting on noise.

Finally, by validating the model out-of-sample, I show that the words selected have predictive power for the future, reducing the possibility that they are artifacts of overfitting.

C. *Risky words and economic uncertainty*

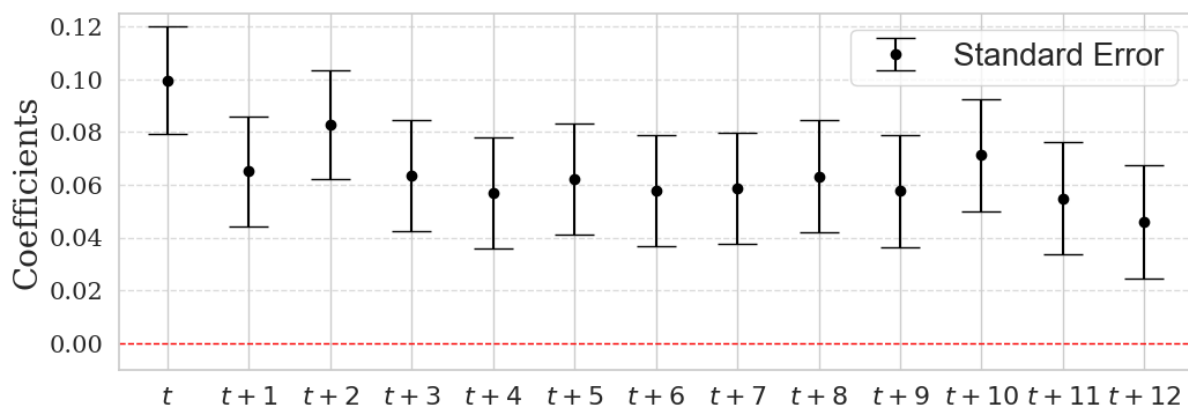
Figure A6 in the Appendix shows the total number of risky words that LASSO picks each month, along with VIX (implied 30-day volatility of the S&P 500 index) and the Economic Policy Uncertainty (EPU) index of Baker et al. (2016). The risky words that have predictive power for the cross-section of stock returns have predictive power for firm-specific and macroeconomic risk conditions. When the model picks many of these words to indicate the sources of systematic risks, it reflects that the economic environment is complex and affected by multifaceted risks. This suggests that market participants and firms are operating in conditions where multiple sources of risk factors are expected to influence returns, leading to heightened uncertainty.

Figure 7 illustrates the predictive relationship between the aggregated number of risky words and future values of two economic uncertainty metrics: the VIX and EPU. Specifically, I plot the coefficients from regressions where the $\log(\text{VIX})$ and $\log(\text{EPU})$ at current and future points $t, t+1, \dots, t+12$ (up to 12 months ahead) are regressed on $\log(\text{aggregated number of risky words}+1)$ at time t .

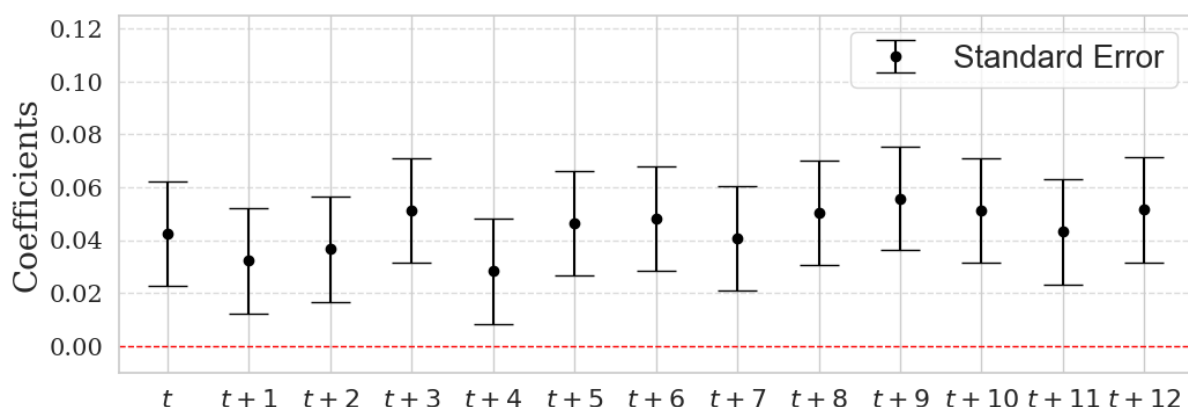
Panel (a) in Figure 7 shows the coefficients of future VIX on the aggregated number of risky words. Each point on the graph represents the estimated coefficient of the number of risky words from the regression of future VIX on the aggregated risky word count at month t . The positive and statistically significant coefficients across the time horizons indicate that an increase in the number of risky words at t is associated with higher values of the VIX in subsequent months, up to 12 months ahead.

The finding suggests that the aggregated number of return-predictive risky words reflects underlying expectations of market participants regarding upcoming periods of heightened uncertainty or volatility. This is consistent with the notion that these words, chosen by LASSO regression for their power to explain stock returns, are indicative of systematic risk factors that investors anticipate will affect market stability. The persistence of positive coefficients up to 12 months implies that these risks are not short-lived but can influence market sentiment and implied volatility over an extended period.

Similarly, panel (b) shows the coefficients of future EPU on aggregated number of risky words. The predictive nature of risky words on EPU implies that these words capture latent information about systematic risks that are priced into market expectations. This latent information can manifest in real-world metrics like VIX and EPU, which respond to the market's



(a) Coefficients of future VIX on aggregated number of risky words



(b) Coefficients of future economic uncertainty on aggregated number of risky words

Figure 7. Risky words predictive power for VIX and Economic Policy Uncertainty (EPU)

This figure depicts the relationship between the aggregated number of risky words in firms' risk sections and future economic uncertainty indicators, specifically the VIX and Economic Policy Uncertainty (EPU). Panel (a) presents the regression coefficients where $\log(\text{VIX})$ at time horizons $t, t+1$, up to $t+12$ months is regressed on $\log(\text{aggregated risky words} + 1)$ at time t . The positive and statistically significant coefficients imply that an increase in the number of risky words at time t is linked to higher future VIX values, extending up to a year. Panel (b) shows the regression results for $\log(\text{EPU})$ with similar time horizons, indicating that more risky words being picked by the model at t corresponds to increased EPU values in the following months. The time period is 2000-2023.

evolving risk perception. The ability of risky words to predict these future conditions underscores their role as a barometer for economic uncertainty, signaling shifts in expected volatility and policy uncertainty.

D. Risky words factor

In this section, I gauge the risky words long-short portfolio against the 13 factors from [Jensen et al. \(2021\)](#), and vice versa. Part (1) in Table 8 shows the alphas and betas of risky words long-short portfolio against each of the 13 factors of 8 in the period 2005-2023. All of the monthly alphas are above 1.10%, and statistically significant, implying that the risky words strategy is not explained by any of these factors. Part (2) reverses the analysis and shows the alpha of each factor when regressed on the risky words long-short portfolio. The results suggest that the risky words portfolio explains most of the time-variation in these 13 factors as most alphas appear

	(1)		(2)		adj R^2
	alpha	beta	alpha	beta	
accruals	1.66*** [3.63]	-1.03** [-2.06]	0.12** [2.03]	-0.02** [-2.06]	0.01
debt issuance	1.52*** [3.32]	0.80 [1.10]	0.04 [0.83]	0.01 [1.10]	0.00
investment	1.60*** [3.50]	-0.40 [-1.48]	0.13 [1.15]	-0.02 [-1.48]	0.01
low leverage	1.56*** [3.40]	-0.02 [-0.07]	0.03 [0.23]	-0.00 [-0.07]	-0.00
low risk	1.48*** [3.40]	0.70*** [4.94]	-0.10 [-0.52]	0.14*** [4.94]	0.09
momentum	1.37*** [3.95]	1.53*** [13.06]	-0.31** [-2.06]	0.28*** [13.06]	0.43
profit growth	1.51*** [3.73]	3.35*** [7.96]	-0.09 [-1.50]	0.07*** [7.96]	0.22
profitability	1.36*** [3.08]	0.93*** [4.24]	0.09 [0.65]	0.08*** [4.24]	0.07
quality	1.10** [2.57]	1.60*** [6.34]	0.14 [1.33]	0.09*** [6.34]	0.15
seasonality	1.44*** [3.15]	1.73** [2.11]	0.05 [1.34]	0.01** [2.11]	0.01
short term reversal	1.63*** [3.59]	-1.01** [-2.36]	0.11 [1.51]	-0.02** [-2.36]	0.02
size	1.49*** [3.51]	-1.26*** [-6.20]	0.12 [0.95]	-0.12*** [-6.20]	0.14
value	1.67*** [3.73]	-0.63*** [-3.55]	0.30* [1.80]	-0.08*** [-3.55]	0.05

Table 8- Testing Asset Pricing factors

This table evaluates the performance of a risky words long-short portfolio relative to 13 factors of [Jensen et al. \(2021\)](#). Part (1) shows the alphas and betas of the risky words portfolio against each factor for the period 2005–2023. Part (2) reverses the analysis, showing the alphas of each factor when regressed on the risky words portfolio. Most alphas are insignificant, suggesting that the risky words portfolio captures a significant portion of the time variation in these factors and offers a better explanation than the market factor. The adjusted R^2 for each regression is reported in the last column.

insignificant. The risky words portfolio provides a better explanation of these factors compared to the market factor. The GRS F -statistic, which tests whether all factors jointly have a zero alpha, decreases from 4.50 (when using the market portfolio as the regressor) to 3.98 (when using the risky words portfolio as the regressor).

VII. Summary and Conclusion

This paper introduces an innovative method to predict stock returns through the textual content of firms' risk disclosures in annual reports. By cross-sectionally running LASSO regressions of returns on the text of risk sections, I identify "risky words": words that have independent predictive power for stock returns. A strategy that times these risky words generates an annual alpha of up to 22% out-of-sample during 2005–2023, and the results do not vanish after controlling for industry, sentiment, firm characteristics, sentiment, and other textual features. Further, by training a word embedding model on the text of risk sections, I cluster these risky words into 14 orthogonal themes, such as energy, healthcare, and manufacturing, which collectively represent systematic risks influencing expected returns.

Key findings highlight that risky words predict returns through dynamic coefficients, reflecting changing economic conditions. My approach consistently identifies a small set of significant predictors, even as the textual dataset grows in size. The aggregated number of risky words predicts broader economic uncertainty indicators, such as the VIX and policy uncertainty indices.

Overall, this paper advances the understanding of systematic risks and their joint connection to stock returns using a novel textual analysis approach. It underscores the dynamic and high-dimensional nature of risk factors, offering a novel perspective that takes advantage of cutting-edge natural language processing techniques. My methodology not only reveals which risks matter most to investors but also provides a framework for decomposing expected returns into their risk-based components.

References

- Baker, Scott R, Nicholas Bloom, and Steven J Davis, 2016, Measuring economic policy uncertainty, *The Quarterly Journal of Economics* 131, 1593–1636.
- Bao, Yang, and Anindya Datta, 2014, Simultaneously discovering and quantifying risk types from textual risk disclosures, *Management Science* 60, 1371–1391.
- Bolton, Patrick, and Marcin Kacperczyk, 2023, Global pricing of carbon-transition risk, *The Journal of Finance* 78, 3677–3754.
- Bryzgalova, Svetlana, Jiantao Huang, and Christian Julliard, 2022, Bayesian solutions for the factor zoo: We just ran two quadrillion models, *The Journal of Finance* .
- Bybee, Leland, Bryan Kelly, Asaf Manela, and Dacheng Xiu, 2024, Business news and business cycles, *The Journal of Finance* 79, 3105–3147.
- Bybee, Leland, Bryan Kelly, and Yinan Su, 2023, Narrative asset pricing: Interpretable systematic risk factors from news text, *The Review of Financial Studies* 36, 4759–4787.
- Campbell, John L, Hsinchun Chen, Dan S Dhaliwal, Hsin-min Lu, and Logan B Steele, 2014, The information content of mandatory risk factor disclosures in corporate filings, *Review of Accounting Studies* 19, 396–455.
- Cao, Sean, T Clifton Green, Lijun Gillian Lei, and Shaojun Zhang, 2023, Expert network calls, *Fisher College of Business Working Paper* 013.
- Carhart, Mark M, 1997, On persistence in mutual fund performance, *The Journal of Finance* 52, 57–82.
- Chen, Andrew Y, and Tom Zimmermann, 2021, Open source cross-sectional asset pricing, *Critical Finance Review*, *Forthcoming* .
- Cohen, Lauren, Christopher Malloy, and Quoc Nguyen, 2020, Lazy prices, *The Journal of Finance* 75, 1371–1415.
- Davis, Steven J, Stephen Hansen, and Cristhian Seminario-Amez, 2020, Firm-level risk exposures and stock returns in the wake of covid-19, Technical report, National Bureau of Economic Research.

- Fama, Eugene F, and Kenneth R French, 1992, The cross-section of expected stock returns, *The Journal of Finance* 47, 427–465.
- Fama, Eugene F, and Kenneth R French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.
- Fama, Eugene F, and Kenneth R French, 2008, Dissecting anomalies, *The Journal of Finance* 63, 1653–1678.
- Fama, Eugene F, and Kenneth R French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.
- Fama, Eugene F, and Kenneth R French, 2016, Dissecting anomalies with a five-factor model, *The Review of Financial Studies* 29, 69–103.
- Fama, Eugene F, and Kenneth R French, 2018, Choosing factors, *Journal of Financial Economics* 128, 234–252.
- Fama, Eugene F, and James D MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of political economy* 81, 607–636.
- Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, 2020, Taming the factor zoo: A test of new factors, *The Journal of Finance* 75, 1327–1370.
- Florackis, Chris, Christodoulos Louca, Roni Michaely, and Michael Weber, 2023, Cybersecurity risk, *The Review of Financial Studies* 36, 351–407.
- Gabaix, Xavier, Ralph SJ Koijen, Robert Richmond, and Motohiro Yogo, 2023, Asset embeddings, *Available at SSRN 4507511* .
- García, Diego, Xiaowen Hu, and Maximilian Rohrer, 2023, The colour of finance words, *Journal of Financial Economics* 147, 525–549.
- Gaulin, Maclean Peter, 2017, *Risk fact or fiction: The information content of risk factor disclosures*, Ph.D. thesis, Rice University Houston, Texas.
- Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri, 2021, Economic predictions with big data: The illusion of sparsity, *Econometrica* 89, 2409–2437.

- Gibbons, Michael R, Stephen A Ross, and Jay Shanken, 1989, A test of the efficiency of a given portfolio, *Econometrica: Journal of the Econometric Society* 1121–1152.
- Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *The Review of Financial Studies* 33, 2223–2273.
- Hanley, Kathleen Weiss, and Gerard Hoberg, 2019, Dynamic interpretation of emerging risks in the financial sector, *The Review of Financial Studies* 32, 4543–4603.
- Harvey, Campbell R, and Yan Liu, 2021, Lucky factors, *Journal of Financial Economics* 141, 413–435.
- Harvey, Campbell R, Yan Liu, and Heqing Zhu, 2016, ... and the cross-section of expected returns, *The Review of Financial Studies* 29, 5–68.
- Hassan, Tarek A, Stephan Hollander, Laurence Van Lent, Markus Schwedeler, and Ahmed Tahoun, 2023a, Firm-level exposure to epidemic diseases: Covid-19, sars, and h1n1, *The Review of Financial Studies* 36, 4919–4964.
- Hassan, Tarek A, Stephan Hollander, Laurence Van Lent, and Ahmed Tahoun, 2019, Firm-level political risk: Measurement and effects, *The Quarterly Journal of Economics* 134, 2135–2202.
- Hassan, Tarek A, Jesse Schreger, Markus Schwedeler, and Ahmed Tahoun, 2023b, Sources and Transmission of Country Risk, *The Review of Economic Studies* rdad080.
- Hassan, Tarek Alexander, Stephan Hollander, Laurence Van Lent, and Ahmed Tahoun, 2023c, The global impact of brexit uncertainty, *The Journal of Finance* .
- Hoberg, Gerard, and Gordon Phillips, 2016, Text-based network industries and endogenous product differentiation, *Journal of political economy* 124, 1423–1465.
- Hoberg, Gerard, and Gordon M Phillips, 2024, Scope, scale and concentration: The 21st century firm, *The Journal of Finance* .
- Hsu, Po-hsuan, Kai Li, and Chi-yang Tsou, 2023, The pollution premium, *The Journal of Finance* 78, 1343–1392.
- Jensen, Theis Ingerslev, Bryan T Kelly, and Lasse Heje Pedersen, 2021, Is there a replication crisis in finance?, *The Journal of Finance, Forthcoming* .

- Kim, Alex, Maximilian Muhn, and Valeri Nikolaev, 2023, From transcripts to insights: Uncovering corporate risks using generative ai, *arXiv preprint arXiv:2310.17721* .
- Koijen, Ralph SJ, Tomas J Philipson, and Harald Uhlig, 2016, Financial health economics, *Econometrica* 84, 195–242.
- Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2018, Interpreting factor models, *The Journal of Finance* 73, 1183–1223.
- Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2020, Shrinking the cross-section, *Journal of Financial Economics* 135, 271–292.
- Li, Feng, 2008, Annual report readability, current earnings, and earnings persistence, *Journal of Accounting and economics* 45, 221–247.
- Li, Qing, Hongyu Shan, Yuehua Tang, and Vincent Yao, 2024, Corporate climate risk: Measurements and responses, *The Review of Financial Studies* hhad094.
- Lopez-Lira, Alejandro, 2021, Why do managers disclose risks accurately? textual analysis, disclosures, and risk exposures, *Economics Letters* 204, 109896.
- Lopez-Lira, Alejandro, 2023, Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns, *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper* .
- Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, *The Journal of Finance* 66, 35–65.
- Loughran, Tim, and Bill McDonald, 2016, Textual analysis in accounting and finance: A survey, *Journal of Accounting Research* 54, 1187–1230.
- Loughran, Tim, and Bill McDonald, 2020, Textual analysis in finance, *Annual Review of Financial Economics* 12, 357–375.
- Neuhierl, Andreas, Otto Randl, Christoph Reschenhofer, and Josef Zechner, 2023, Timing the factor zoo, *Available at SSRN* 4376898 .
- Patton, Andrew J, and Michela Verardo, 2012, Does beta move with news? firm-specific information flows and learning about profitability, *The Review of Financial Studies* 25, 2789–2839.

Sautner, Zacharias, Laurence Van Lent, Grigory Vilkov, and Ruishen Zhang, 2023, Firm-level climate change exposure, *The Journal of Finance* 78, 1449–1498.

Savor, Pavel, and Mungo Wilson, 2016, Earnings announcements and systematic risk, *The Journal of Finance* 71, 83–138.

Seyfi, Sina, 2022, Neighbouring assets, *Available at SSRN 4311284* .

Seyfi, Sina, 2023, Essence of the cross section, *Available at SSRN* .

Seyfi, Sina, 2024, Basis portfolios, *Available at SSRN* .

Appendices

A. LASSO Regression

The objective function is as follows:

$$\text{minimize} \left(\frac{1}{2n_t} \sum_{i=1}^{n_t} \left(r_{i,t+1} - a_t - \sum_{j=1}^{n_t} b_{j,t} \text{word}_{i,j,t} \right)^2 + \lambda_t \sum_{j=1}^{N_t} |b_{j,t}| \right),$$

at each month t , n_t is the number of unique words (unigrams and bigrams) available, and λ_t is the regularization parameter at time t that controls the sparsity degree in the coefficient estimates.

For each value of λ_t , the LASSO regression model is trained on the 4 training folds. The performance of the model is then evaluated on the validation fold that was left out during training. The key metric I use to measure performance is the Mean Squared Error (MSE), which quantifies the difference between the observed and predicted values. The MSE is calculated for each of the 5 validation sets, and the results are averaged to produce a single performance measure for that λ_t value.

After testing various λ_t values, the one that yields the lowest average MSE is selected as the optimal λ_t for that month. This method not only helps in achieving a balance between the model's complexity and its predictive accuracy but also aids in avoiding overfitting, ensuring that the model performs consistently well on new, unseen data. This cross-validation process is carried out each month, ensuring that each monthly model is optimized for that specific period.

B. Word Embeddings

The Continuous Bag of Words (CBOW) model operates in a way that, for a given target word W_k , the model takes a context window size $c = 5$ and uses the words $W_{k-c}, \dots, W_{k-1}, W_{k+1}, \dots, W_{k+c}$ to predict W_k . The model optimizes a function that maximizes the likelihood of the target word given its context.

Training CBOW involves optimizing an objective function to maximize the probability of predicting a target word given its context words. In contrast to Skip-Gram, which predicts context words from a target word, CBOW predicts a target word from surrounding context words.

The objective function for the CBOW model is defined as follows:

$$\max \sum_{k=1}^K \log P(W_k | W_{k-c}, \dots, W_{k-1}, W_{k+1}, \dots, W_{k+c})$$

Where K is the total number of words in the corpus, c is the size of the context window (number of words considered to the left and right of the target word), and

$$P(W_k | W_{k-c}, \dots, W_{k-1}, W_{k+1}, \dots, W_{k+c})$$

is the probability of predicting the target word W_k given its context words.

For each target word W_k , CBOW considers a window of words c on either side (e.g., if $c = 5$, it uses five words before and two words after the target word as the context). The model takes these context words and aggregates them (typically by averaging their vector representations) to create a combined context representation. The combined context representation is fed into the hidden layer of a simple neural network. The model predicts the target word W_k using a softmax function, which outputs the probability distribution over the entire vocabulary. The training objective is to maximize the likelihood of correctly predicting the target word given the context. The optimization minimizes the negative log-likelihood of the target words using the stochastic gradient descent technique.

The probability of predicting the target word W_k given the context words is:

$$P(W_k | W_{k-c}, \dots, W_{k-1}, W_{k+1}, \dots, W_{k+c}) = \frac{e^{x(W_k) \cdot \bar{x}(\text{context})}}{\sum_{W=1}^{|V|} e^{x(W) \cdot \bar{x}(\text{context})}}$$

Where $x(W_k)$ is the vector representation of the target word in the output layer, $\bar{x}(\text{context})$ is the average of the vectors representing the context words and $|V|$ is the size of the vocabulary.

The training Process involves the following steps:

1. Forward Pass: For each training sample, the input layer collects the vectors of context words. These vectors are averaged to produce a context vector $\bar{x}(\text{context})$. The context vector is passed through the hidden layer to predict the target word.
2. Loss Calculation: The loss function used is the negative log-likelihood of the correct target word:

$$L = -\log P(W_k | W_{k-c}, \dots, W_{k-1}, W_{k+1}, \dots, W_{k+c}).$$

3. Backpropagation: The error is backpropagated through the network to update the word vectors (weights) in the embedding matrix and the output layer.
4. Optimization Techniques: Negative sampling.

In natural language processing tasks with large vocabularies, computing the full softmax is slow because it requires updating weights for every word in the vocabulary during each training step, which can be computationally prohibitive. Negative sampling addresses this issue by simplifying the training process. Instead of updating weights for all words, it updates weights only for the correct (positive) word-context pairs and a small number of randomly chosen incorrect (negative) word pairs.

The training process with negative sampling involves maximizing the probability of a target word given its context while simultaneously minimizing the probability of a few negative words that do not co-occur with the target. These negative words are selected based on their frequency distribution, with more common words being more likely to be chosen. This approach significantly reduces training time because the computation scales with the number of negative samples rather than the size of the entire vocabulary. The number of negative samples is a hyperparameter, and I set it to 5, balancing between approximation accuracy and computational efficiency. Negative sampling enables training on large corpora with better speed while still providing a reliable approximation of the full softmax distribution.

Overall, the CBOW model offers a powerful and efficient way to train word embeddings by predicting a target word from its context words. It captures general semantic meanings effectively, making it suitable for many NLP applications. While it may not capture rare word representations as well as Skip-Gram, it excels in scenarios where the focus is on frequent word predictions and when computational efficiency is paramount.

C. Risky Clusters

Below is the list of all clusters as well as all risky words that get a non-zero coefficient in the risk sections. The numbers in parenthesis show how many times the word has received a non-zero coefficient in 288 cross-sectional regressions. The words that are pinpointed by an underline are the main risky words and the core of each cluster. The words with a positive average correlation are shown in green, and those with a negative average coefficient are displayed in red.

1. Energy

- oil (173), gas (152), steel (79), energy (69), coal (64), fuel (61), oil natural (52), natural gas (49), oil gas (49), power (37), electric (34), crude (28), raw (27), travel (27), solar (25), utility (21), freight (18), smaller reporting (17), segment (15), gold (13), plants (13), paper (12), nuclear (11), prices (9), generation (8), transportation (8), labor (8), renewable (7), pipelines (7), forward (7), utilities (6), crude oil (6), wind (6), emissions (5), agricultural (5), metals (5), natural (5), raw materials (5), electricity (4), cost (4), increased (4), time (3), increase (3), fuels (3), regulated (3), chemical (3), metal (3), expect (3), intermediate (2), climate (2), driving (2), industrial (2), aluminum (2), commodity (2), gathering (2), petroleum (2), minerals (2), transition (2), weight (2), greenhouse (2), machines (2), pronouncements (1), driver (1), truck (1), continue going (1), carbon (1), diesel (1), catastrophe (1), served (1), resource (1), minimum wage (1), selling prices (1), resources (1), engines (1), line (1), index (1), north (1), pass (1), domestic (1), chemicals (1), availability (1), price common (1), changes rates (1), currently (1), sustain (1), imports (1), estimate (1), floor (1), west (1)

2. Drug

- clinical (162), products (161), product candidates (106), fda (97), drug (94), patent (50), product (48), nasdaq (45), equipment (42), trials (41), candidates (37), vehicles (36), government (36), device (26), pharmaceutical (26), intellectual (23), additional (23), table contents (22), regulatory (22), collaborators (22), regulations (19), tests (19), programs (17), marketing (15), table (14), diagnostic (13), devices (13), patents (13), generic (13), laws (13), clinical trials (12), trial (11), able (11), cells (10), develop (10), therapeutic (9), product candidate (9), candidate (9), factors (8), libor (8), clearance (8), program (8), drug candidates (7), phase (7), s government (7), license (7), cancer (7), risk factors (7), commercial (6), clinical studies (6), approval (6), therapy (6), pipeline (6), drugs (6), committee (6), future product (5), report (5), marketing approval (5),

application (5), obtain (5), proprietary (5), sfas (5), vaccine (4), preclinical (4), potential (4), patient (4), court (4), gene (3), studies (3), agent (3), looking statements (3), reference (3), epa (3), annual report (3), trademarks (3), product liability (3), sec (3), procedure (3), registration statement (3), nda (2), candidates develop (2), nonclinical (2), products technologies (2), disease (2), terms (2), achieve (2), compensation committee (2), study (2), item management (2), clinical trial (2), collaboration (2), new products (2), agencies (2), safety (2), direct marketing (2), treatment (2), unable (1), pre clinical (1), pcaob (1), rule (1), products product (1), medicinal (1), commercialize (1), approvals (1), delays (1), s performance (1), covid (1), therapeutics (1), designation (1), approved (1), gaap (1), future products (1), field (1), day (1), cms (1), safe (1), investigational (1), monitoring (1), comments (1), formulations (1), licensors (1), report form (1), feasibility (1), requirements able (1), technological (1), legislation (1), sec s (1), s products (1), licensing (1), performance (1), exclusive (1), acceptance (1), market acceptance (1), governmental (1), limited (1), process (1), registration (1), oral (1), statements contained (1), discovery (1), diseases (1), head (1), filed (1)

3. Company

- company (142), stock (129), company s (63), warrants (58), mr (52), notes (48), units (44), shares (38), common stock (37), common (36), companies (25), reporting company (25), securities (25), class common (23), corporation s (19), management (17), corporation (16), dividends (15), common shares (15), company required (15), stockholders (14), class (14), shareholders (14), reporting (13), merger (13), nyse (13), dr (9), preferred stock (9), director (9), series (8), group (8), quarter (8), registrant (8), reporting companies (7), convertible notes (6), trading (5), independent (5), executive (4), company smaller (4), combination (4), shares common (4), ownership change (4), senior notes (3), company defined (3), s common (3), controls (3), class b (2), combined company (2), ordinary shares (2), preferred shares (2), holders (2), addition (2), change (2), share (2), entities (2), enterprises (2), incorporated (2), concern (2), ag (2), b common (1), elected (1), shares class (1), consummate (1), impact company (1), redemption (1), officer (1), holding company (1), investment company (1), stockholder (1), rule securities (1), maryland (1), able raise (1), ordinary (1), companies required (1), image (1), groups (1), preferred securities (1), delisted (1), raise (1), community (1), effect company (1), decline (1), corp (1), stock fluctuate (1), independence (1), statement (1), direct (1), fiscal company (1), statement form (1), company subsidiaries (1), factors company (1)

4. Loan

- loans (139), bank (136), mortgage (87), loan (80), credit (62), agreement (50), real estate (39), capital (38), debt (33), portfolio (31), risk (31), investment (30), contracts (29), banking (28), shall (27), credit facility (25), loan losses (25), indebtedness (23), real (23), deposits (21), subsidiaries (14), investments (14), agreements (14), statements (14), acquisitions (14), estate (14), iv (13), financing (13), banks (12), federal (11), forward looking (11), venture (10), incorporated reference (10), servicing (9), institutions

(9), senior (9), credit agreement (8), liquidity (8), ratings (8), initial (7), transactions (7), university (7), restated (6), automobile (6), lender (6), fdic (6), covenants (6), warrant (6), access (6), charter (5), default (5), revolving credit (4), penny (4), note (4), broker (4), institution (4), rating (4), funding (4), exhibit (4), senior secured (3), page (3), purchaser (3), reflect (3), asset (2), accounts (2), securitization (2), loan portfolio (2), forum (2), defined rule (2), secured credit (2), lenders (2), bonds (2), family (2), brokerage (2), federal government (2), underwriting (2), allowance loan (2), s treasury (2), bank s (2), guaranty (2), treasury (2), receivables (2), borrower (2), city (2), term (2), debentures (2), affiliates (2), ppp (1), ppp loan (1), news (1), nasdaq capital (1), loan facility (1), borrowing base (1), amended restated (1), term loan (1), master (1), maturities (1), credit facilities (1), employer (1), arrangements (1), non performing (1), capitalized (1), near term (1), amended (1), provision loan (1), amendment (1), credit markets (1), lending (1), collateral (1), lines (1), financial covenants (1), cautionary (1), reverse (1), rent (1), backed (1), portfolios (1), mortgages (1), alliance (1), line credit (1), sound (1), settlement (1), paragraph (1), express (1), affiliate (1)

5. Properties

- stores (129), properties (119), merchandise (108), homes (65), tenants (41), facility (34), retail (33), vehicle (31), construction (28), reserves (28), brands (28), title (26), home (26), environmental (22), programming (18), assets (17), store (17), intellectual property (15), inventory (15), brand (14), facilities (14), land (14), franchise (11), contents (11), rental (10), retailers (10), lease (10), search (10), restaurants (9), property (9), traffic (9), backlog (8), centers (8), stations (7), communities (7), residential (7), channel (7), orders (6), shopping (5), school (5), space (4), florida (4), bid price (4), associates (4), leases (4), housing (4), new york (4), market area (4), southern (3), point (2), wholesale (2), locations (2), tenant (2), ownership (2), temporary (2), york (2), staff (2), building (2), volume (2), warehouse (2), sites (2), trucks (1), living (1), permits (1), clinics (1), proposed transaction (1), leads (1), amazon (1), data centers (1), restructure (1), unresolved (1), weeks (1), bulk (1), membership (1), dollar (1), land use (1), residents (1), branches (1), specialty (1), force (1), profitable (1), options (1), station (1), items (1), branded (1), children (1), trademark (1), unit (1), current assets (1), premises (1)

6. Clients

- customers (112), clients (96), patients (31), licensees (20), distributors (19), customer (17), partners (15), dealers (15), suppliers (15), manager (14), client (14), users (12), drivers (9), contract (7), consumers (6), members (5), funds (5), providers (4), vendors (3), employees (3), channel partners (2), contractor (2), distributor (2), brokers (2), subcontractors (2), professionals (2), merchant (1), investors (1), advisor (1), ordering (1), parties (1), meet (1), strategic partners (1), resellers (1), advisory (1), capital market (1), participating (1), end users (1), competitors (1), officers (1), cause (1), service providers (1), confidential (1)

7. Software

- **software** (106), **services** (93), **wireless** (82), **solutions** (66), **internet** (60), **advertising** (59), **digital** (58), **network** (51), **consumer** (47), **technology** (47), **content** (44), **platform** (43), **television** (39), **video** (36), **technologies** (35), **online** (33), **new** (32), **systems** (32), **data** (26), **mobile** (25), **communications** (25), **cable** (24), **service** (23), **information** (22), **networks** (22), **optical** (22), **carriers** (19), **applications** (17), **web** (17), **card** (16), **radio** (15), **com** (15), **media** (13), **provide information** (12), **products services** (10), **enterprise** (10), **voice** (10), **industry** (9), **websites** (9), **use** (8), **e commerce** (8), **design** (7), **risks related** (7), **networking** (7), **satellite** (7), **server** (7), **transmission** (6), **fleet** (6), **subscriber** (6), **emerging growth** (5), **financial services** (5), **open source** (5), **assurance** (5), **fiber** (5), **storage** (5), **tm** (5), **fcc** (5), **commerce** (4), **subscription** (4), **electronic** (4), **domain** (4), **risks relating** (4), **broadband** (4), **imaging** (4), **training** (4), **pcs** (4), **internal control** (3), **spectrum** (3), **cloud** (3), **charging** (3), **carrier** (3), **information required** (3), **continue** (3), **infrastructure** (3), **designs** (3), **operators** (3), **model** (3), **platforms** (3), **solution** (3), **plan** (3), **auction** (3), **check** (3), **cards** (3), **surgical** (3), **document** (3), **bandwidth** (3), **ip** (2), **subscriptions** (2), **hosting** (2), **computing** (2), **offerings** (2), **regulation s** (2), **rail** (2), **information technology** (2), **drive** (2), **audio** (2), **distance** (2), **learning** (1), **personal data** (1), **smart** (1), **solutions services** (1), **emerging** (1), **servers** (1), **virtual** (1), **google** (1), **cyber** (1), **connectivity** (1), **advice** (1), **services certain** (1), **services new** (1), **engineers** (1), **website** (1), **billing** (1), **communication** (1), **technology systems** (1), **models** (1), **continued listing** (1), **initiatives** (1), **personal** (1), **regulation** (1), **marketplace** (1), **superior** (1), **mail** (1), **telephone** (1), **timely** (1), **competitive** (1), **internal** (1), **representatives** (1), **core** (1), **defense** (1), **based** (1), **detection** (1), **outsourced** (1), **interactive** (1), **professional services** (1), **library** (1), **measurement** (1), **modular** (1), **pc** (1), **support** (1), **phone** (1), **web site** (1), **databases** (1), **long distance** (1), **computer** (1), **integrated** (1), **software services** (1), **offer** (1), **web sites** (1), **mission** (1), **fax** (1), **speed** (1)

8. Business

- **china** (90), **semiconductor** (90), **business** (75), **operations** (69), **financial** (59), **business combination** (46), **results** (44), **smaller** (42), **future** (39), **market** (37), **foreign** (31), **chinese** (31), **automotive** (25), **adversely** (23), **components** (21), **operating** (19), **material** (18), **adverse** (15), **risks** (14), **global** (12), **markets** (12), **rights** (12), **growth** (11), **prospects** (10), **operating results** (10), **currency** (9), **business financial** (8), **joint** (8), **results operations** (8), **ability** (8), **value** (8), **international** (8), **oem** (8), **condition results** (7), **united** (7), **countries** (7), **collaborative** (7), **conditions** (6), **economic** (6), **significant** (6), **condition** (6), **entertainment** (6), **business prospects** (5), **eu** (5), **financial reporting** (5), **american** (5), **financial condition** (5), **litigation** (5), **businesses** (5), **india** (5), **mexico** (5), **pandemic** (4), **canada** (4), **electronics** (4), **ventures** (4), **oems** (4), **control financial** (3), **flows** (3), **materially** (3), **industries** (3), **foreign exchange** (3), **coast** (3), **profitability** (3), **business operating** (3), **silicon** (3), **laser** (3), **operations cash** (2), **united states** (2), **science** (2), **material adverse** (2), **condition operating** (2), **financial position** (2), **strategic** (2), **joint venture** (2), **activities** (2), **south** (2), **processes** (2), **pacific** (2), **biotechnology** (2), **negative** (2), **discussion** (2), **national** (2), **circuit**

(2), display (2), extraordinary (2), s business (2), covid global (1), taiwan (1), global pandemic (1), relating operations (1), operations prospects (1), factors described (1), america (1), japan (1), financial results (1), jurisdictions (1), restrictions (1), financial performance (1), coronavirus (1), canadian (1), uk (1), effect business (1), condition liquidity (1), trade (1), retain existing (1), cause results (1), way (1), led (1), tier (1), weakness (1), respond (1), s discussion (1), sales operating (1), results financial (1), proceedings (1), actions (1), currencies (1), position (1), compete (1), export (1), negatively (1), similarly (1), market price (1), central (1), harm business (1), relationships (1), cyclical (1), business plan (1), workforce (1), operating history (1), play (1), ms (1), financial statements (1), expand (1), number (1), light (1), future results (1), impacted (1), circuits (1), memory (1)

9. Applicable

- **applicable** (88), required (38), impact (36), changes (29), certain (19), affect (15), required provide (14), listing (12), subject (7), exchange (6), obligations (5), section (4), effect (2), required make (2), contractual arrangements (2), requirements (2), proposed (2), qualify (2), related (2), act required (2), required include (2), agents (1), topic (1), described item (1), non (1), related intellectual (1), local state (1), changes local (1), federal regulations (1), adverse effect (1), consecutive (1), sarbanes oxley (1), publish (1), compliant (1), local (1), relating (1), later (1), assure (1), information relating (1), contents item (1), change control (1), contained (1)

10. Insurance

- **insurance** (83), reimbursement (19), claims (15), payors (8), warranty (7), public (5), life insurance (5), compensation (5), licenses (4), liability (3), insurance company (2), health insurance (1), payers (1), frequency (1), casualty (1), workers (1), repair (1), bargaining (1), insured (1)

11. Drilling

- **exploration** (83), **drilling** (81), research (33), project (29), offshore (14), projects (13), fracturing (13), going concern (11), hydraulic (10), mineral (8), president (6), wells (5), restructuring (5), research development (5), acquisition (4), period (4), mining (3), hydraulic fracturing (3), control (3), substantial (2), hedging (2), include disclosure (1), exploration production (1), suspension (1), seismic (1), guarantee (1), expansion (1), verification (1), start (1)

12. Healthcare

- **food** (75), **healthcare** (66), **water** (60), care (53), medical (44), medicare (41), health (28), hospitals (23), education (22), item (21), including (14), security (13), materials (12), physicians (11), party (9), packaging (9), ingredients (8), medicaid (8), waste (7), blood (6), vice (5), weather (5), california (5), health care (4), wage (4), act (4), site (4), air (4), tissue (3), pharmacy (3), events (3), vice president (3), orphan (2), quality (2), prescription (2), affiliated (2), emergency (2), looking (2), physician (2), texas (2), remedy (2), aspect (2), animal (1), nevada (1), climate change (1), aca (1), public

offering (1), supplies (1), oxley (1), legal (1), human (1), managed care (1), private (1), compliance (1), senior vice (1)

13. Manufacturing

- **production** (73), **manufacturing** (69), sales (54), **development** (53), costs (31), **distribution** (17), **cell** (17), demand (13), manufacturers (12), test (11), plant (10), **result** (10), **processing** (7), **laboratory** (5), **launch** (4), shipping (4), capacity (4), **supply** (4), **testing** (3), **separation** (3), **assembly** (2), **refining** (2), **personnel** (2), **life** (2), **transit** (1), logistics (1), **precision** (1), **fulfillment** (1), **contract manufacturers** (1), **port** (1), **sourcing** (1), **supplier** (1), **recovery** (1), **closure** (1), **manufactured** (1), **chain** (1), **integration** (1), **engineering** (1)

14. Income

- **income** (70), **trust** (62), **million** (60), **reit** (59), tax (53), revenue (48), general partner (41), **partnership** (39), **revenues** (32), **fiscal** (32), **partner** (31), rates (29), **cash** (24), **december** (23), **pension** (18), **preferred** (14), **rate** (14), distributions (13), **convertible** (12), **llc** (12), **harm** (12), **net sales** (11), **ended** (11), **series preferred** (10), **price** (10), consolidated (10), **earnings** (10), **year** (10), **net** (9), **state** (9), **directors** (9), **dated** (8), allowance (8), **royalty** (8), **losses** (8), **goodwill** (7), **march** (7), **states** (5), **holdings** (5), **net loss** (5), **spin** (5), **ended december** (5), **bankruptcy** (5), **impairment** (5), **fiscal year** (5), **cash flows** (4), **loss** (4), **expenses** (4), **revolving** (4), **accounting** (4), **approximately** (4), **september** (4), **ebitda** (3), **annual** (3), **liabilities** (3), **taxable** (3), **fees** (3), **average** (3), **purchase** (2), **date** (2), **resident** (2), **june** (2), **percent** (2), **condition cash** (2), **tax assets** (2), **employee** (2), **minimum** (2), **deferred** (2), **deposit** (2), **plus** (2), **included** (2), **participant** (2), **san** (2), **savings** (2), **registered trademark** (2), **pay** (1), **foreign tax** (1), **ii** (1), **indenture** (1), **repurchase** (1), **federal income** (1), **salary** (1), **milestone** (1), **fee** (1), **april** (1), **fund** (1), **years** (1), **liquidation** (1), **proceeds** (1), **approximately million** (1), **sale** (1), **called** (1), **excise** (1), **net revenues** (1), **intangible** (1), **basis** (1), **results cash** (1), **profits** (1), **billion** (1), **year ended** (1), **margins** (1), **october** (1), **issued** (1), **compared** (1), **derivative** (1), **duty** (1), **costs expenses** (1), **ended march** (1), **ended september** (1), **quarter ended** (1), **respectively** (1), **fiscal years** (1), **believes** (1), **closing** (1), **income loss** (1), **months ended** (1), **total revenues** (1), **reserve** (1), **course** (1), **virginia** (1)

D. Figures

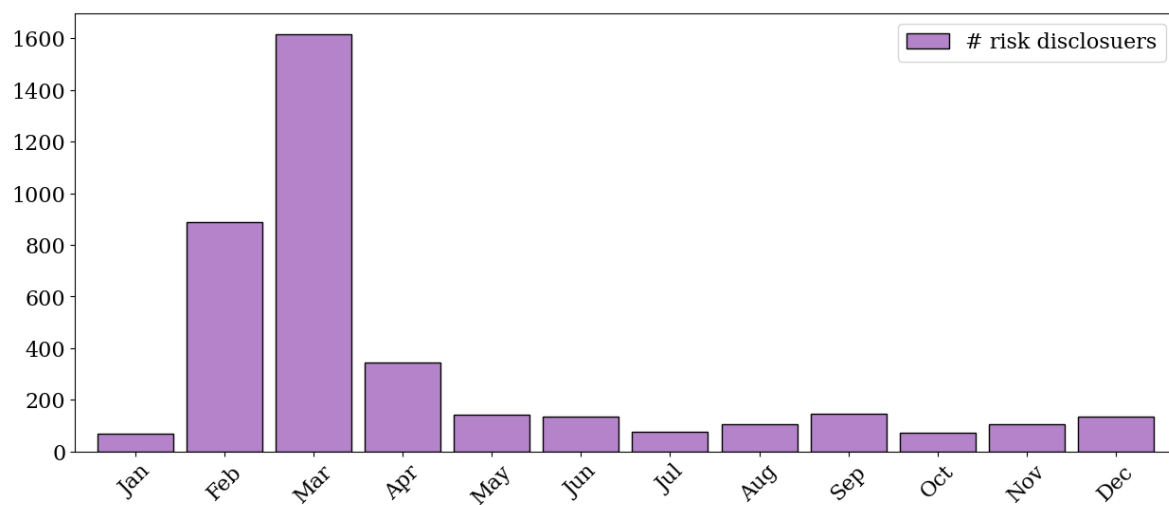
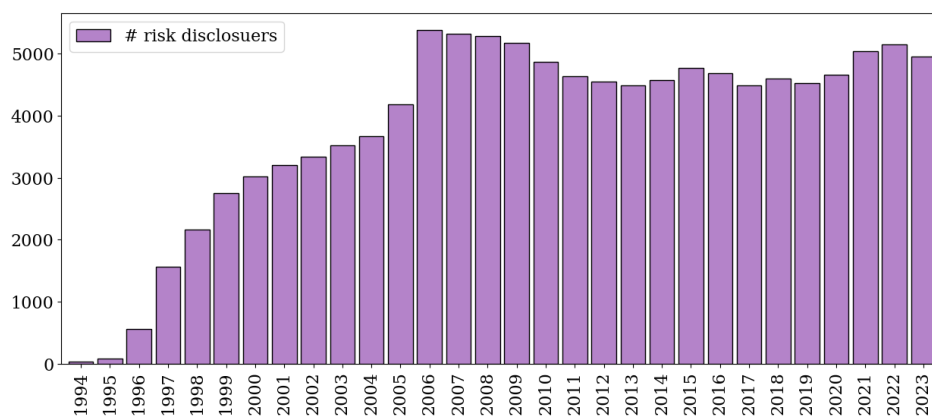
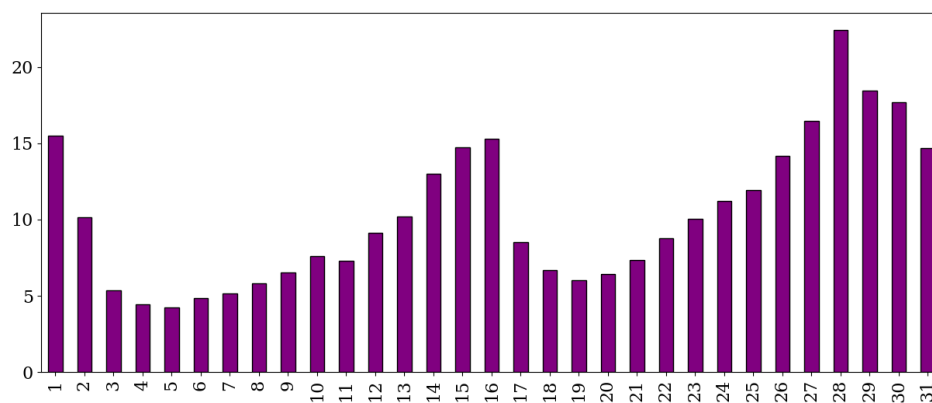


Figure A1. Average number of reports per month

This figure shows the average number of 10-K reports per month between 1995-2023. As most companies fiscal year is at the end of the calendar year, more than two-thirds of reports are filed in February and March. The number of reports per year and the average number of reports per day are shown in Figure A2, panels (a) and (b).



(a) The number of risk sections filed each year



(b) The average number of files per day in each month

Figure A2. Risk sections statistics

Panel (a) shows the number of files that contain a risk section per year. Panel (b) shows the average number of files per day in each month. The average number of files per month is shown in Figure A1.

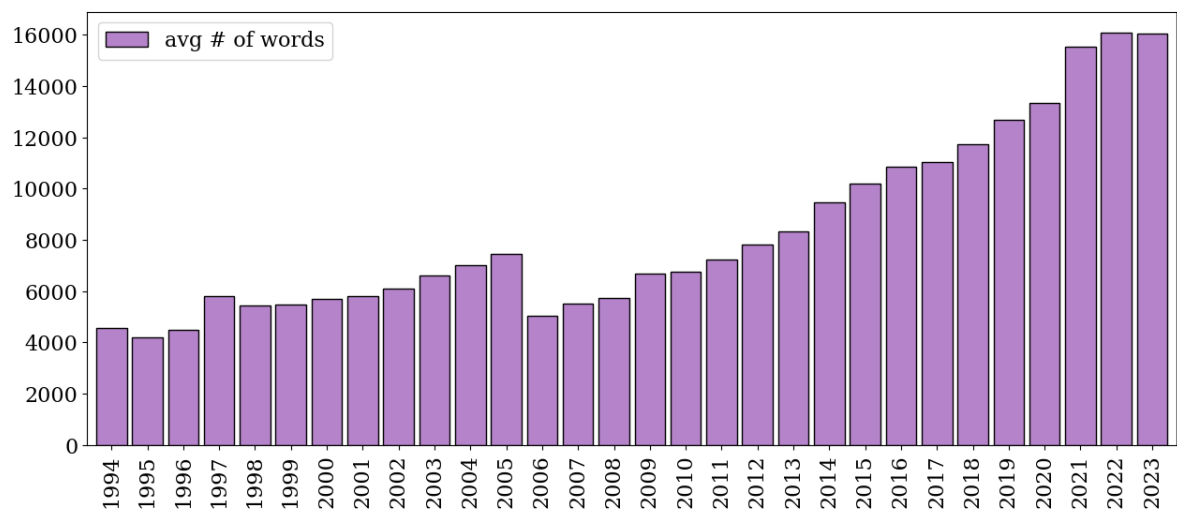


Figure A3. The average length of risk sections per year

This Figure displays the average word count of the "Risk Factors" sections annually. In 2006, a year after the implementation of this requirement, the average length was about 5,000 words. By 2023, this figure had increased threefold, reaching approximately 16,000 words.

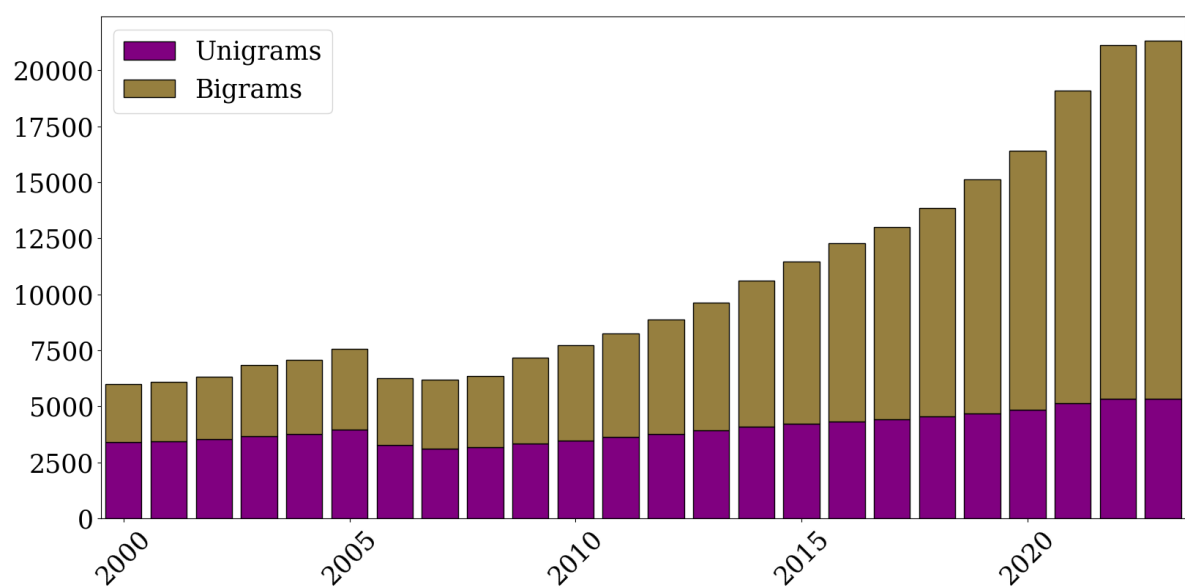


Figure A4. Average number of unigrams and bigrams in the risk sections

This figure shows the average number of unigrams and bigrams in the risk sections in each year. When creating the TF-IDF matrix, I exclude unigrams and bigrams that are either too frequent (appearing in more than 95% of disclosures during that month) or too infrequent (appearing in less than 5% of risk sections in that month).

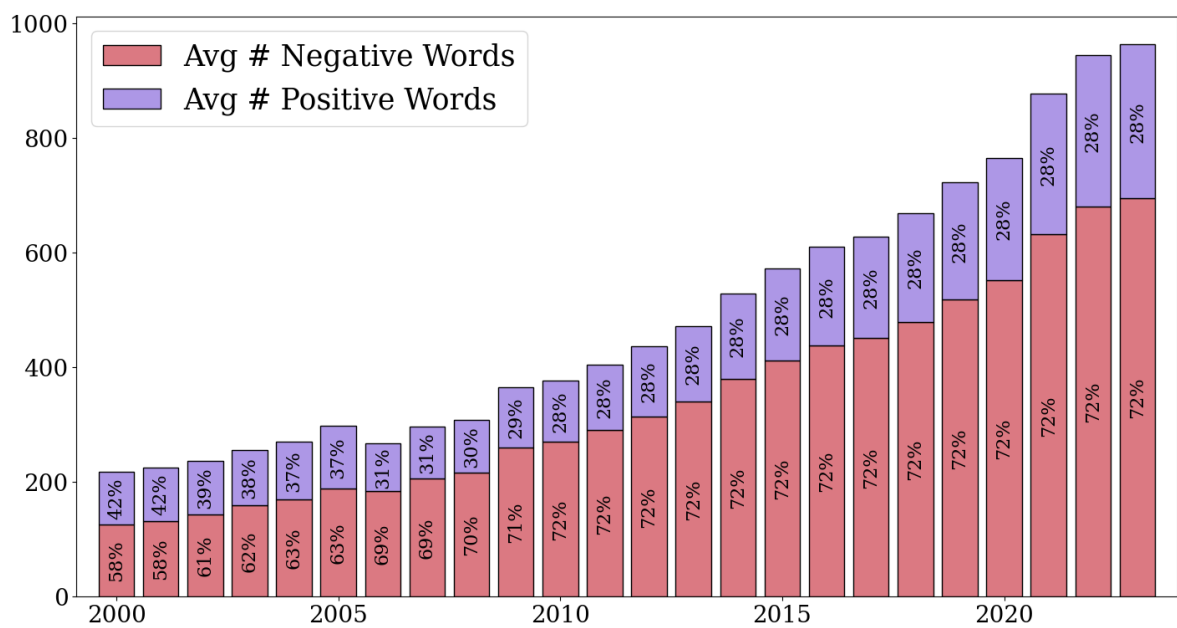


Figure A5. The average number of positive and negative words in risk sections per year

This figure illustrates the annual average counts of positive (blue) versus negative (red) words within the risk sections, consistently revealing a higher frequency of negative words each year. The positive and negative words are calculated based on the traditional dictionary of [Loughran and McDonald \(2011\)](#). The corresponding figure based on the Machine Learning dictionary of [García et al. \(2023\)](#) is shown in figure 1.

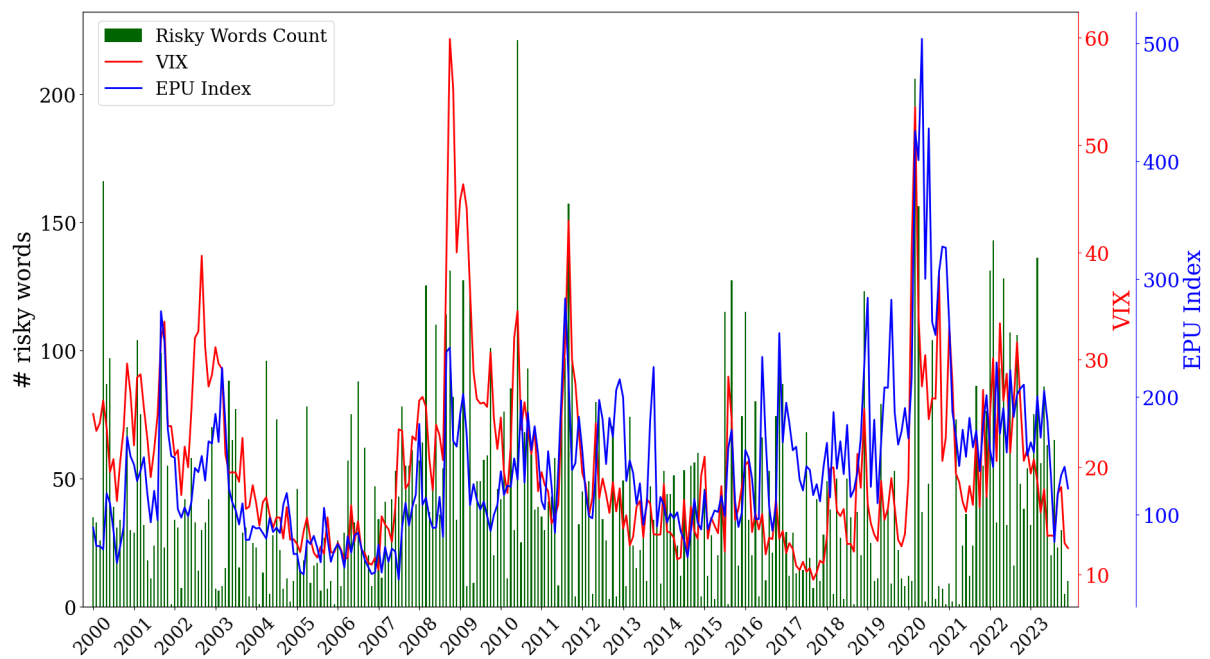


Figure A6. Aggregate number of risky words, VIX and EPU

This figure illustrates the monthly count of risky words identified by LASSO, (on the left axis) plotted alongside the VIX (which represents the implied 30-day volatility of the S&P 500 index) and the Economic Policy Uncertainty (EPU) index on the right axis.

	Energy	Drug	Company	Loan	Properties	Clients	Software	Business	Applicable	Insurance	Drilling	Healthcare	Manufacturing	Income
Energy	24.28													
Drug	0.33	14.78												
Company	0.19	-0.82	21.71											
Loan	-0.02	-0.65	3.60	15.20										
Properties	4.74	1.87	1.36	3.07	17.38									
Clients	-0.04	1.11	-0.99	0.96	5.74	32.12								
Software	2.46	2.41	-1.89	-1.11	2.78	4.30	15.89							
Business	2.00	0.25	0.69	-0.75	0.34	0.77	0.51	9.17						
Applicable	-3.13	-1.15	-1.37	-0.66	-4.84	-2.72	-3.28	-1.56	20.63					
Insurance	1.19	0.85	-1.76	2.25	1.96	0.76	0.00	-1.43	-0.80	42.40				
Drilling	8.77	3.50	-1.08	0.59	3.87	-1.30	0.67	1.81	-3.25	0.40	34.04			
Healthcare	4.81	1.97	-1.58	-0.77	2.52	1.80	2.95	-0.82	-1.24	4.12	0.92	16.84		
Manufacturing	4.55	3.70	-3.97	-2.56	2.23	0.99	1.15	2.10	-2.24	-1.25	4.55	1.05	22.74	
Income	-0.16	-1.76	3.99	3.59	1.12	-1.55	-1.97	-0.01	-0.91	-0.04	0.24	-0.98	-0.25	10.61

Table A1- The average cosine similarity between clusters

The table reports the average cosine similarity between clusters, showing how semantically similar or dissimilar the words are within and across clusters. The diagonal values indicate the average similarity within each cluster, revealing high semantic coherence among words. For example, the *energy* cluster has an average similarity of 24.28%. The off-diagonal values represent the similarities between different clusters, generally reflecting lower relationships. The clusters are shown in Figures 2 and 3.

	mean				α_{FF8}			
	1	2	3	LS	1	2	3	LS
Durbl	-0.18 [-0.25]	0.61 [1.14]	1.75** [2.60]	1.92*** [3.52]	-1.04*** [-2.93]	-0.24 [-0.96]	0.84* [1.94]	1.88*** [3.59]
Enrgy	0.51 [0.57]	0.57 [0.67]	0.99 [1.37]	0.48 [1.17]	-0.64 [-0.97]	-0.53 [-0.85]	0.08 [0.15]	0.72* [1.78]
HiTec	0.56 [1.13]	0.92** [2.10]	1.17*** [2.77]	0.61*** [3.19]	-0.10 [-0.58]	0.25* [1.75]	0.49*** [3.52]	0.59*** [3.44]
Hlth	0.57 [1.08]	0.66 [1.38]	1.00** [2.00]	0.43 [1.43]	0.20 [0.72]	0.30 [1.38]	0.61** [2.18]	0.41 [1.31]
Manuf	0.45 [0.85]	1.07** [2.38]	1.28*** [2.85]	0.84*** [3.43]	-0.37** [-2.21]	0.16 [1.24]	0.40** [2.16]	0.77*** [3.97]
NoDur	0.26 [0.54]	0.78** [2.08]	0.89** [2.52]	0.63** [2.30]	-0.45* [-1.85]	0.06 [0.34]	0.22 [1.30]	0.67*** [2.82]
Other	0.35 [0.83]	0.71* [1.91]	0.89** [2.50]	0.54** [2.46]	-0.11 [-0.69]	0.11 [1.29]	0.27*** [2.76]	0.38** [2.06]
Shops	0.29 [0.55]	1.04** [2.25]	1.14** [2.56]	0.85*** [3.30]	-0.55** [-2.47]	0.11 [0.58]	0.25 [1.31]	0.80*** [3.54]
Telcm	0.27 [0.44]	0.47 [1.09]	1.23** [2.31]	0.97** [2.25]	-0.43 [-1.12]	-0.20 [-0.75]	0.55 [1.52]	0.98** [2.33]
Utils	0.60* [1.85]	0.84*** [3.04]	0.87*** [3.18]	0.28 [1.11]	-0.09 [-0.40]	0.27 [1.24]	0.29 [1.38]	0.38 [1.61]

Table A2- Sorting stocks within each industry

I report monthly average excess returns (left side) and eight-factor alphas (right side) from the Fama and French model for portfolios sorted based on the risky words strategy and each of the Fama-French 10 industries. From January 2005 to December 2023, I sort stocks at the end of each month in each industry in ascending order based on their predicted value from the risky words model and allocate them into three groups. The LS is a 3-minus-1 long-short portfolio. The portfolios are equally weighted and I rebalance portfolios monthly. The t -statistics are reported in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.