Risk Factors that Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns

Latest version here

Alejandro Lopez-Lira*

The Wharton School, University of Pennsylvania

First Version: November 2018

This Version: November 2019

Abstract

I exploit unsupervised machine learning and natural language processing techniques to elicit the risk factors that firms themselves identify in their annual reports. I quantify the firms' exposure to each identified risk, design an econometric test to classify them as either systematic or idiosyncratic, and construct factor mimicking portfolios that proxy for each undiversifiable source of risk. The portfolios are priced in the cross-section and contain information above and beyond the commonly used multi-factor representations. A model that uses only firm identified risk factors (FIRFs) performs at least as well as traditional factor models, despite not using any information from past prices or returns.

*I am grateful to Jules van Binsbergen, Joao Gomes, Nick Roussanov, Itamar Drechsler, Winston Dou, Amir Yaron, Lars Hansen, Seth Pruitt, Thomas Sargent, Alexander Belyakov, Marco Grotteria, Andrew Wu (discussant), Ryan Israelsen (discussant), Simona Abis (discussant), Michael Halling (discussant), Rujian Chen (discussant), Chen Xue (discussant), Jose Penalva (discussant), Diego Garcia (discussant), Jose Luis Montiel Olea (discussant), Kyre Lahtinen (discussant), as well as participants from numerous conferences for helpful discussions and feedback. I am grateful and acknowledge the financial support provided by the Mack Institute for Innovation Management, the Rodney L. White Center for Financial Research, and The Jacobs Levy Equity Management Center for Quantitative Financial Research. Please email me for comments or suggestions: joselop@wharton.upenn.edu.

1 Introduction

> "The best hope for finding pricing factors that are robust out of sample and across different markets, is to try to understand the fundamental macroeconomic sources of risk." Cochrane (2005)

Since the introduction of the Capital Asset Pricing Model (CAPM) over fifty years ago, the asset pricing literature has witnessed a tremendous growth in potential additional factors that could help explain the cross-section of expected stock returns. The formidable economic challenge comes in finding and interpreting economically relevant risk factors. I propose a novel approach to this challenge by eliciting the risk factors that firms themselves identify in their annual reports. I then evaluate which ones are systematic, which ones are priced, and whether or not they contain information above and beyond the standard factors and characteristics in the literature.[1]

To accomplish this, I use machine learning to identify the risks that firms face by applying textual analysis techniques on their annual reports. Then, I design an econometric test to classify them as either systematic or idiosyncratic. Furthermore, I show they contain information beyond the usual factors. Finally, I provide a model that uses only firm identified risk factors (FIRFs), that performs at least as well as traditional models, while being economically motivated, literally described by words and not using any information from past prices.

How to get a list of the fundamental risks in the economy? Firms are required to disclose extensively every risk they face in their annual reports. Figure 1 shows an excerpt from a specific section of Apple's 10-K annual report: Item 1A Risk Factors. The actual document is much longer, and while it contains extremely detailed information, it is impractical to read every single page of every company's report every single year. Figure 2 shows the result of applying machine learning: we get a representation of the document as the proportion of the

1. As an example of additional factors see Fama and French (1992), Fama and French (1993), Fama and French (2015), Hou, Xue, and Zhang (2015), and Stambaugh and Yuan (2017) among many, many others

risk disclosure that the company allocates to each risk (risk weights). We also obtain each of the common risks companies discuss, as Figure 8 shows. Figure 3 shows the International Risk (topic): each risk is described by (a distribution over) words. The most discussed risks are Innovation, Demand, Production, International, and Property Risk.[2]

Are the risks captured by the machine learning algorithm idiosyncratic or systematic? In theory, just because many companies are discussing a particular risk, say, a possible shortage in the supply of computer parts, it does not mean it cannot be diversified away by investors. In practice, an international war can trigger a shortage in the supply of components, which seems systematic.

I solve the problem of classifying risks into systematic and idiosyncratic or diversifiable by designing an econometric test. The intuition for the test is the following: only the systematic components appear in the covariances between companies. Hence, if two companies are more exposed to a systematic risk factor, their covariance will increase, whereas if those companies are more exposed to a diversifiable risk, it will not. For the 2006-2019 period, International, China, Oil, and Credit Risk are the most systematic ones in the sense that they increase the covariance between stocks the most.[3]

Are these risk priced and well described by traditional models? I use the risk weights to understand the impact of the risks in the first and second moments of returns by running Fama-Macbeth cross-sectional regressions and time-series regressions, in addition to regressions on correlations to assess which risks are systematic. A Gibbons, Ross, and Shanken (1989) (GRS) test shows statistically significant evidence that the Fama-French Five-Factor model does not span all of the risks. The unexplained portion of the returns of the risks is usually referred to as 'α_s', although the term here is slightly misleading since we are considering portfolios that track the risks that firms face.

2. The machine learning algorithm is called Latent Dirichlet Allocation (LDA), Blei, Ng, and Jordan (2003). Note that while it is reasonable to have concerns about the reliability of the risk disclosures, there is ample evidence in the accounting literature that shows risk disclosures are truthful and informative, e.g., Campbell et al. (2014), Gaulin (2019).

3. Hanley and Hoberg (2019) propose a different test to asses when risks in the financial sector become systematic.

Can we get interpretable factors that represent economic risk? How much can we explain with the common risks faced by the firms? I construct factor-mimicking portfolios for each specific risk and form a factor model using the most discussed risks. I test the capacity of these factors to price the cross-section of returns using the set of 25 Book-to-Market, and 49 Industry Portfolios available from Kenneth French's website.[4]

With the LDA algorithm, we get a set of 25 risks. For parsimony and comparability with existing factor models, I focus on four risks to form the factor model in the main analysis.[5] I select the risks that companies spend more time discussing at the beginning of the sample to avoid any look-ahead-bias concern. Note that I do not use any information about returns to select the factors, neither from the test set (the 49 industry portfolios, the 25 book-to-market portfolios, and the anomalies) nor from the 25 potential risk factors. [6]

Despite the factor model being constructed to capture the different sources of risks for the firms and especially for interpretability, the model performs surprisingly well. The factor model has a statistical fit at least as good as the leading models in the literature: the factor models of Fama and French (2015), Stambaugh and Yuan (2017), and Hou, Xue, and Zhang (2015). Nevertheless, it is essential to reiterate that the objective of the paper is not to run a horse race with the current models.[7]

For example, using the GRS test,(Gibbons, Ross, and Shanken (1989)) which tests the null of no-mispricing ($\alpha_i = 0$), and where lower values of the GRS statistic correspond to lower evidence of mispricing (and higher p-values): with the set of 49 industry portfolios,

4. I choose these portfolios as the test set following the critique of Lewellen, Nagel, and Shanken (2010). I include additional tests using the profitability-investment portfolios and the anomalies portfolios from Stambaugh and Yuan (2017)

5. I also perform an extensive analysis, including using all of the portfolios, LASSO regression (Tibshirani (1996)) for dimensionality reduction, clustering the portfolios using the covariance matrix as in Stambaugh and Yuan (2017), and clustering the companies using the disclosed risks, the results are similar and available in the online Appendix.

6. See Section 9 for details

7. Cochrane (2005): "Thus, it is probably not a good idea to evaluate economically interesting models with statistical horse races against models that use portfolio returns as factors. Economically interesting models, even if true and perfectly measured, will just equal the performance of their own factor-mimicking portfolios, even in large samples. They will always lose in sample against ad-hoc factor models that find nearly ex-post efficient portfolios."

the GRS statistic is .88, with a corresponding p-value of 68% which means we cannot reject the no-mispricing null; compare to the GRS statistic of 1.55 for the Fama and French (2015) model with a p-value of 4.5% in which we can reject the no-mispricing null.[8]

The paper continues as follows: Section 2 provides the literature review; Section 3 describes the data sets and addresses concerns about the reliability of the annual reports; Section 4 describes extensively the process to recover risks from the annual reports; Section 5 describes the risks; Section 6 describes the test to assess whether a risk is systematic; Section 7 shows the results of running cross-sectional regressions; Section 8 describes the portfolio formation and whether the risks contain novel information; Section 9 describes the performance of the risks as factors; and Section 10 concludes.

8. Additionaly, it succeeds in explaining a large fraction of the time-series variation of the cross-section of returns (measured by an average R^2 of 63 %, comparable to the 68% average R^2 obtained with the Fama and French (2015) Model). However, Lewellen, Nagel, and Shanken (2010) advise against using R^2 to compare between models. See Section 8 for details.

Figure 1: Excerpt from Item 1A: Risk Factors in Apple Inc. Annual Report

Item 1A. Risk Factors

The following discussion of risk factors contains forward-looking statements. These risk factors may be important to understanding other statements in this Form 10-K. The following information should be read in conjunction with Part II, Item 7, "Management's Discussion and Analysis of Financial Condition and Results of Operations" and the consolidated financial statements and related notes in Part II, Item 8, "Financial Statements and Supplementary Data" of this Form 10-K.

The business, financial condition and operating results of the Company can be affected by a number of factors, whether currently known or unknown, including but not limited to those described below, any one or more of which could, directly or indirectly, cause the Company's actual financial condition and operating results to vary materially from past, or from anticipated future, financial condition and operating results. Any of these factors, in whole or in part, could materially and adversely affect the Company's business, financial condition, operating results and stock price.

Because of the following factors, as well as other factors affecting the Company's financial condition and operating results, past financial performance should not be considered to be a reliable indicator of future performance, and investors should not use historical trends to anticipate results or trends in future periods.

Global and regional economic conditions could materially adversely affect the Company.

The Company's operations and performance depend significantly on global and regional economic conditions. Uncertainty about global and regional economic conditions poses a risk as consumers and businesses may postpone spending in response to tighter credit, higher unemployment, financial market volatility, government austerity programs, negative financial news, declines in income or asset values and/or other factors. These worldwide and regional economic conditions could have a material adverse effect on demand for the Company's products and services. Demand also could differ materially from the Company's expectations as a result of currency fluctuations because the Company generally raises prices on goods and services sold outside the U.S. to correspond with the effect of a strengthening of the U.S. dollar. Other factors that could influence worldwide or regional demand include changes in fuel and other energy costs, conditions in the real estate and mortgage markets, unemployment, labor and healthcare costs, access to credit, consumer confidence and other macroeconomic factors affecting consumer spending behavior. These and other economic factors could materially adversely affect demand for the Company's products and services.

In the event of financial turmoil affecting the banking system and financial markets, additional consolidation of the financial services industry, or significant financial service institution failures, there could be tightening in the credit markets, low liquidity and extreme volatility in fixed income, credit, currency and equity markets. This could have a number of effects on the Company's business, including the insolvency or financial instability of outsourcing partners or suppliers or their inability to obtain credit to finance development and/or manufacture products resulting in product delays; inability of customers, including channel partners, to obtain credit to finance purchases of the Company's products; failure of derivative counterparties and other financial institutions; and restrictions on the Company's ability to issue new debt. Other income and expense also could vary materially from expectations depending on gains or losses realized on the sale or exchange of financial instruments; impairment charges resulting from revaluations of debt and equity securities and other investments; changes in interest rates; increases or decreases in cash balances; volatility in foreign exchange rates; and changes in fair value of derivative instruments. Increased volatility in the financial markets and overall economic uncertainty would increase the risk of the actual amounts realized in the future on the Company's financial instruments differing significantly from the fair values currently assigned to them.

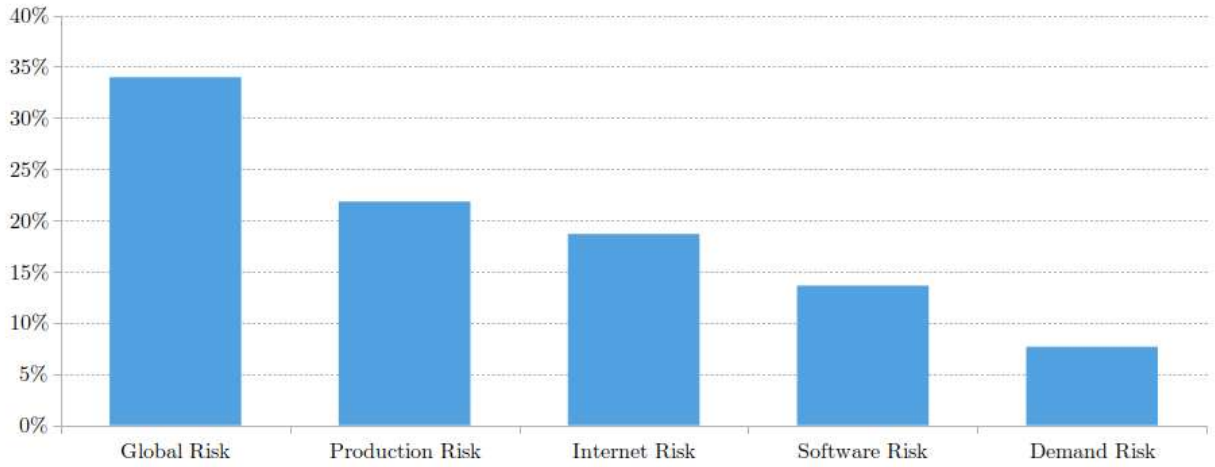
Global markets for the Company's products and services are highly competitive and subject to rapid technological change, and the Company may be unable to compete effectively in these markets.

The Company's products and services compete in highly competitive global markets characterized by aggressive price cutting and resulting downward pressure on gross margins, frequent introduction of new products, short product life cycles, evolving industry standards, continual improvement in product price/performance characteristics, rapid adoption of technological and product advancements by competitors and price sensitivity on the part of consumers.

The Company's ability to compete successfully depends heavily on its ability to ensure a continuing and timely introduction of innovative new products, services and technologies to the marketplace. The Company believes it is unique in that it designs and develops nearly the entire solution for its products, including the hardware, operating system, numerous software applications and related services. As a result, the Company must make significant investments in R&D. The Company currently holds a significant number of patents and copyrights and has registered and/or has applied to register numerous patents, trademarks and service marks. In contrast, many of the Company's competitors seek to compete primarily through aggressive pricing and very low cost structures, and emulating the Company's products and infringing on its intellectual property. If the Company is unable to continue to develop and sell innovative new products with attractive margins or if competitors infringe on the Company's intellectual property, the Company's ability to maintain a competitive advantage could be adversely affected.

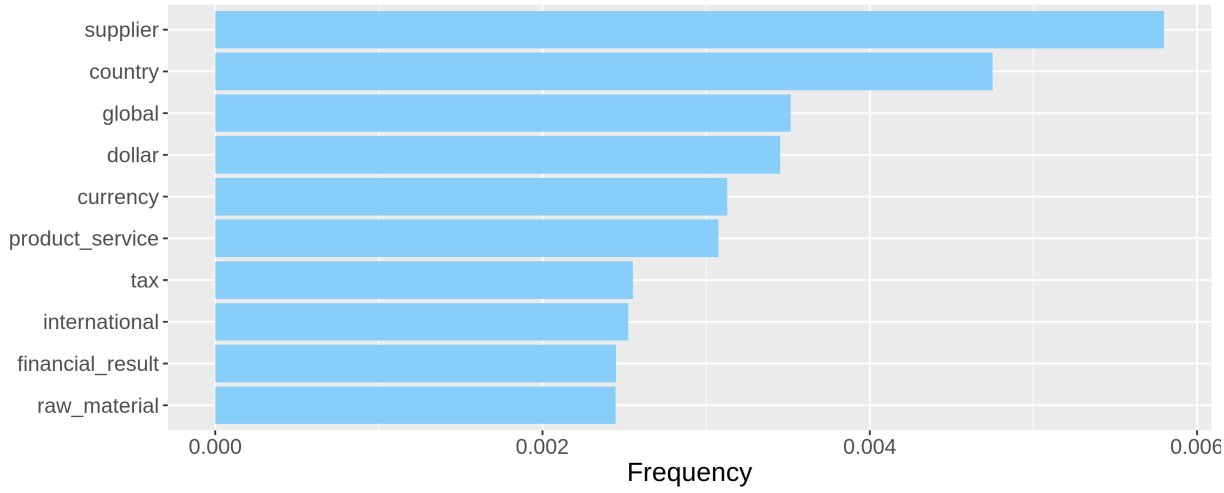
The Figure shows the first page out of ten of Item 1A: Risk Factors in Apple Inc. 10-K 2016 annual report. The document is available on the SEC EDGAR database.

Figure 2: Percentage of the risk disclosure that Apple Inc. allocates to each risk



The table shows the percentages of the risk disclosure that Apple Inc. allocates to each type of risk in the Section 1A: Risk Factors for their 2016 annual report. The table only shows the five most discussed risks. The values are obtained using Latent Dirichlet Allocation. See Sections 3 and 5 for details

Figure 3: International Risk Topic



Distribution of the 10 most frequent words for the International Risk Topic (excluding 'company')

2 Related Literature

My paper makes contributions in two different branches of literature: (1) machine learning and text analysis in finance, and (2) cross-sectional asset pricing.

I contribute to the recent strand of the literature that employs text analysis to study a variety of finance research questions (e.g., Jegadeesh and Wu (2013), Campbell et al. (2014), Hoberg and Phillips (2016), Gaulin (2019), Baker, Bloom, and Davis (2016), Ke, Kelly, and Xiu (2019), Bybee et al. (2019), Ke, Montiel Olea, and Nesbit (2019)). See Loughran and McDonald (2016) for an excellent review. Some papers employ text analysis to study a specific risk that the researchers have in mind (e.g. Hassan et al. (2019) for political risk; Loughran, McDonald, and Pragidis (2019) for oil risk). I instead, do not specify any risk ex-ante and instead let them arise naturally from the data using machine learning methods.

The early literature of topic modeling in finance studies the interaction between disclosed risks, volatility, and betas, abstaining from studying the pricing of the disclosed risks due to the short time horizon. Israelsen (2014) is one of the first papers in finance that uses topic modeling on the risk disclosures and focuses on the interaction between several disclosed risks, stock-return volatility, and betas of the Fama-French Four-Factor model in the period 2006-2011, using weekly returns. Bao and Datta (2014) explore the interaction between disclosed risk and volatility to showcase their novel topic modeling technique. Israelsen (2014) and Bao and Datta (2014) use topic modeling for the full period, so the risk weights they use suffer from look-ahead-bias, which I avoid by using an online version of the topic modeling algorithm. Hanley and Hoberg (2018) propose a different way to deal with look-ahead-bias an apply the technique to understand emerging risks in the financial sector, although they abstract from asset pricing implications.

By using risk weights with no look-ahead-bias and a dataset with a significantly longer time horizon, I answer a completely different set of asset pricing questions compared to the previous literature that uses topic modeling: Which of the fundamental risks in the economy are systematic? Are they priced? Are they summarized well by existing models? Can we

get interpretable factors that represent economic risk? How much can we explain with the common risks faced by the firms?

To answer these questions, I design a novel econometric test to distinguish between systematic and idiosyncratic risks and show that they contain information above-and-beyond what is commonly found in the literature. Furthermore, I show that a model that uses only firm identified risk factors performs at least as well as traditional factor models, despite not using any information from past prices or returns.

My paper is of course related to the large literature on cross-sectional stock returns (see, e.g., Cochrane (1991); Berk, Green, and Naik (1999); Gomes, Kogan, and Zhang (2003); Nagel (2005); Zhang (2005); Livdan, Sapriza, and Zhang (2009); Eisfeldt and Papanikolaou (2013); Kogan and Papanikolaou (2014)). See Harvey, Liu, and Zhu (2016) for a recent systematic survey. However, to the best of my knowledge, this is the first paper to propose a test to distinguish between systematic and idiosyncratic risks, characterize which risks are priced, and construct a factor model using the risks disclosed by the firms.

The factor model I form using the firms' disclosed risks complements the literature in the following ways. First, regarding statistical factor models: while they provide an outstanding statistical fit, they are not designed to be interpretable, so naturally it is hard to understand the economics of these factors and whether they represent risk; are generated by behavioral patterns; or represent market inefficiencies, whereas by design, the factors constructed from the firms' risk disclosures represent economic risk.[9]

Second, regarding empirical factor models: while they succeed in explaining empirically puzzling portfolios (portfolios with $\alpha \neq 0$), they usually do so by iteratively adding (some of) the existing anomalies as risk factors.[10] However adding previously discovered anomalies as risk factors naturally generates too many factors, what has been referred as a "factor zoo" (Cochrane (2011)), and disentangling the true risk factors from the anomalies is a

9. See Kelly, Pruitt, and Su (2018), Kozak, Nagel, and Santosh (2018)

10. See for example Fama and French (1992), Fama and French (1993), Fama and French (2015), Hou, Xue, and Zhang (2015), and Stambaugh and Yuan (2017) among many, many others

complicated endeavor.[11] To complicate things further, there are important concerns as to which of these anomalies are significant out-of-sample (Harvey, Liu, and Zhu (2016), McLean and Pontiff (2016)), so adding them as risk factors is at best, risky. Since, by construction, all of the factors in the paper, represent risk, it suffices to identify which of these factors are priced and what assets we can price.

Finally, regarding economic theory models: we know from Merton (1973) that the risk premia of every asset depends on the covariances of the firms' cash-flows with the market wealth and other state variables that affect the stochastic discount factor (SDF). Any characteristic of the firms that makes their dividends covary with either wealth or state variables would affect returns. Asking researchers to identify most of these variables seems like an unworkable task. Firms, however, have a much better understanding of the risks they are facing. Hence, understanding which risks firms face can provide guidance on how to improve our theoretical models.

3 Data

I use three sources of data: the 10-Ks Annual Reports, Compustat, and CRSP.

3.1 10-K Annual Reports

Firms disclose in their annual reports the types of risk they are facing. There can be some concerns about how true and informative these disclosures are, however, there exists ample evidence that the risk disclosures are, indeed, useful and informative.

First, firms are legally required to discuss "the most significant factors that make the company speculative or risky" (Regulation S–K, Item 105(c), SEC 2005) in a specific section of the 10-K annual reports (Section 1A) and could face legal action if they fail to obey the regulation, as well as being vulnerable to lawsuits from investors.

11. Feng, Giglio, and Xiu (2017) however, provide some hope to succeed in this endeavor.

Additionally, Campbell et al. (2014) find that "the type of risk the firm faces determines whether it devotes a greater portion of its disclosures towards describing that risk type... managers provide risk factor disclosures that meaningfully reflect the risks they face and the disclosures appear to be... specific and useful to investors".

In a more recent study, Gaulin (2019) finds that "managers time their identification of new risk factors and removal of previously identified ones to align with the expected occurrence of future adverse outcomes...[and] firms respond to investor demand in a manner consistent with the litigation shield hypothesis... inconsistent with concerns of uninformative boilerplate or 'copy and paste' disclosure".

Finally, all of the annual reports are audited, and it is stated in the General Accepted Accounting Principles that any material information about the risks that the company faces has to be revealed.

I extract the textual risk factors in Section 1A (mandatory since 2005) of each 10-K Annual Report. I collect the 10-Ks from 2005 to 2019 from the EDGAR database on the SEC's website. The 10-Ks come in many different file formats (.txt, .xml, and .html) and have different formatting, so it is quite challenging to automatically extract the Section 1A-Risk Factors, from the 10-K forms. To do so, I first detect and remove the markup language and then use regular expressions with predefined heuristic rules. I end up with a data set consisting of 79304 documents.

To illustrate the kind of disclosures that firms make, consider the excerpt from Apple Inc.'s 2010 10-K annual report below. I incorporate suggested labels regarding the type of risk, and highlight possible key words in red. Note that both labels and key words are just for illustrative purposes, and there is no need to manually label the risks in the paper or define the keywords, since the risks will arise naturally using the LDA algorithm.

- Currency Risk: Demand ... could differ ... since the Company generally raises prices on goods and services sold outside the U.S. to offset the effect of the strengthening of the U.S. dollar change.

- Supplier Risk: The Company uses some custom components that are not common to the rest of the personal computer, mobile communication and consumer electronics industries.

- Competition Risk: Due to the highly volatile and competitive nature of the personal computer, mobile communication and consumer electronics industries, the Company must continually introduce new products

3.2 CRSP and Compustat

I follow the usual conventions regarding CRSP and Compustat data. I focus on monthly returns since the disclosures are done annually. For the accounting and return data, I use the merged CRSP/Compustat database. I use annual firm-level balance sheet data from Compustat due to concerns about seasonality and precision; and monthly returns from CRSP. I use data from the same period as the one where 10-Ks are available: 2006-2019, although not all variables are available for every period.

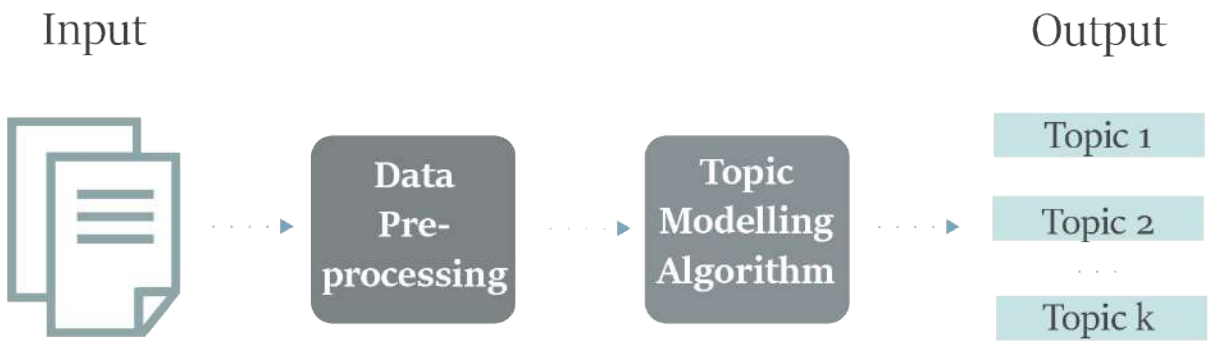
I exclude from the main analysis firms in industries with SIC codes corresponding to the financial industry (SIC in [6000, 7000]). The Five Factors of Fama and French (2015), the momentum factor, and the one-month Treasury-bill rate come from the French data library on Ken French's website. The Stambaugh and Yuan (2017) factors come from their website. The q-factors of Hou, Xue, and Zhang (2015) come from their website.

4 Text Processing

The main takeaway from this section is that we can use machine learning (LDA) to get two objects: the risks that firms are discussing (risk topics) and (2) how much time each company discusses each risk (risk weights). We can get both the risks topics and the risk weights in real time and hence any strategy that bets on specific risks is tradable.

The risks topics are technically called topics in the natural language processing literature

Figure 4: Steps for topic modelling



and formally they are distributions over words. Intuitively, the documents are projected in the risk topic space. Each document is represented by a distribution over topics, the risk weights.

The remainder of the Section is completely optional.

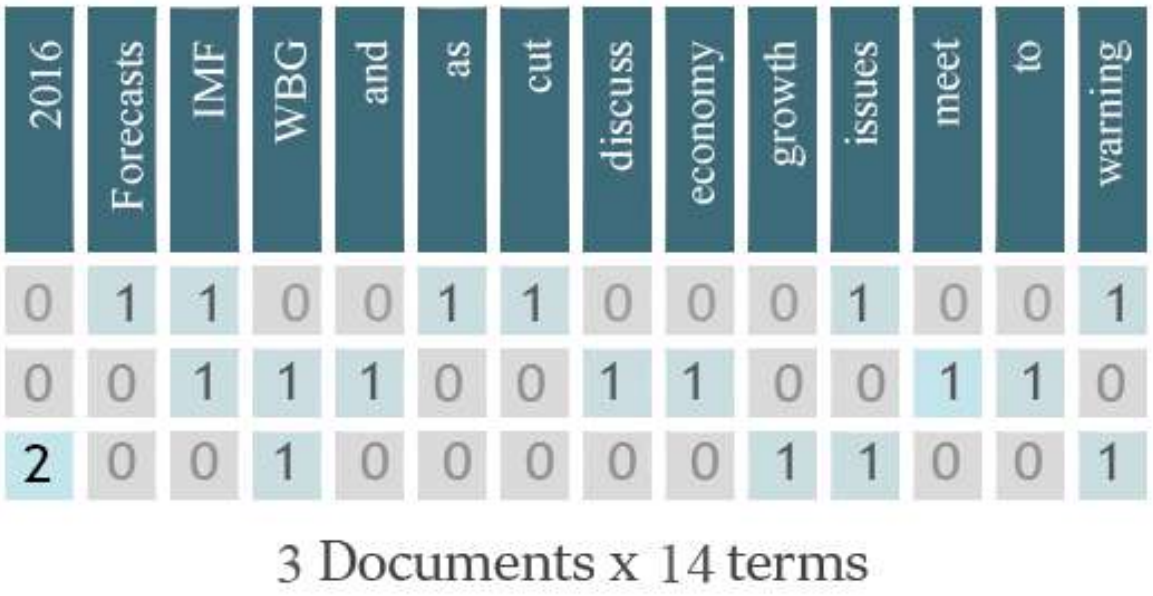
4.1 Bag of Words and Document Term Matrix

We need a way to represent text data for statistical purposes. The Bag of Words model achieves this task. Bag of Words considers a text as a list of distinct words in a document and a word count for each word,[12] which implies that each document is represented as a fixed-length vector with length equal to the vocabulary size. Each dimension of this vector corresponds to the count or occurrence of a word in a document. Traditionally, all words are lowercased to reduce the dimension in half.

It is called a "bag" of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document. Notice that since we only consider the count, the order of the words is lost. When we consider several documents at a time, we end up with a Document Term Matrix (DTM), see Figure 5 for a simplified example. The

12. Manning, Raghavan, and Schütze (2008)

Figure 5: Example of a very simple document term matrix



DTM is typically highly dimensional ($>$ 10,000 columns), since we consider the space of all words used across all documents; it is also very sparse, since typically documents do not use the whole English vocabulary. Because of the huge dimension of the space, we need a dimensionality reduction technique, such as LDA.

Another subtle disadvantage of the Bag of Words model, is that it breaks multi-word concepts such as "real estate" into "real" and "estate", which have to be rejoined later, since counting those words separately will produce different results than counting the multi-word concept.

4.2 Preprocessing

It is common to preprocess the raw text in several steps in order to make the topics more interpretable and to reduce the dimension. The purpose is to reduce the vocabulary to a set of terms that are most likely to reveal the underlying content of interest, and thereby facilitate the estimation of more semantically meaningful topics.

I remove common English words ("the", "and", "or", etc.) and additional terms that do not convey any meaning or are considered legal warnings in the 10-K ("materially adverse", "no assurance", etc.) in order to extract only risks from the text. See the appendix for a full list and a detailed explanation.

Some words represent the same underlying concept. For example, "copy", "copied", and "copying"; all deal with either a thing made to be similar or identical to another or to make a similar or identical version of. The model might treat them differently, so I strip such words to their core. We can achieve this by either stemming or lemmatization, which are fundamental text processing methods for text in the English language.

Stemming helps to create groups of words that have similar meanings and works based on a set of rules, such as remove "ing" at the ends of words.[13] Different types of stemmers are available in standard text processing software such as NLTK (Loper and Bird (2002)), and within the stemmers there are different versions such as PorterStemmer, LancasterStemmer and SnowballStemmer. The disadvantages of stemming is that it cannot relate words that have different forms based on grammatical constructs, for example: "is", "am", and "be" all come from the same root verb, "to be", but stemming cannot prune them to their common form. Another example: the word "better" should be resolved to good, but stemmers would fail to do that. With stemming, there is lot of ambiguity that may cause several different concepts to appear related. For example, "axes" is both a plural form of "axe" and "axis". By chopping of the "s", there is no way to distinguish between the two.

Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form, (Manning, Raghavan, and Schütze (2008)). In order to relate different inflectional forms to their common base form, it uses a knowledge base called WordNet. With the use of this knowledge base, lemmatization can convert words that have a different form and cannot be solved by stemmers, for example converting "are" to "be". The disadvantages of

13. Manning, Raghavan, and Schütze (2008)

lemmatization are that it is slower compared to stemming, however, I use lemmatization to preserve meaning and make the topics more understandable.

Phrase Modeling is another useful technique whose purpose is to (re)learn combinations of tokens that together represent meaningful multi-word concepts. We can develop phrase models by looking for words that co-occur (i.e., appear one after another) together much more frequently than you would expect them to by random chance. The formula to determine whether two tokens A and B constitute a phrase is:

$\frac{count(A,B)-count_{min}}{count(A)*count(B)} * N \geq threshold$, where:

- $count(A)$ is the number of times token A appears in the corpus
- $count(B)$ is the number of times token B appears in the corpus
- $count(A, B)$ is the number of times the tokens A and B appear in the corpus consecutively
- N is the total size of the corpus vocabulary
- $count_{min}$ is a parameter to ensure that accepted phrases occur a minimum number of times
- $threshold$ is a parameter to control how strong of a relationship between two tokens the model requires before accepting them as a phrase

With phrase modeling, named entities will become phrases in the model (so new york would become new_york). We also would expect multi-word expressions that represent common concepts, but are not named entities (such as real estate) to also become phrases in the model.

4.3 Dictionary methods

The most common approach to text analysis in economics relies on dictionary methods, in which the researcher defines a set of words of interest and then computes their counts or

frequencies across documents. However, this method has the disadvantage of subjectivity from the researcher perspective, since someone has to pick the words. Furthermore, it is very hard to get the full list of words related to one concept and the dictionary methods assume the same importance or weight for every word. Since the purpose of the paper is to extract the risks that managers consider important with minimum researcher input, dictionary methods are unsatisfactory.

Furthermore, dictionary methods have other disadvantages, as noted by Hansen, McMahon, and Prat (2018):

> For example, to measure economic activity, we might construct a word list which includes "growth". But clearly other words are also used to discuss activity, and choosing these involves numerous subjective judgments. More subtly, "growth" is also used in other contexts, such as in describing wage growth as a factor in inflationary pressures, and accounting for context with dictionary methods is practically very difficult.

For the purpose of studying the cross-section of returns, the problem is similar to picking which characteristics are important for the returns. The dictionary methods would be equivalent to manually picking which characteristics would enter a regression. The following algorithm, Topic Modelling, is akin to automatic selection methods, such as LASSO (Tibshirani (1996)).

4.4 Topic Models

A topic model is a type of statistical model for discovering a set of topics that describe a collection of documents based on the statistics of the words in each document, and the percentage that each document allocates to each topic. Since in this case, the documents are the risk disclosures from the annual statements and they only concern risks, the topics discovered will correspond to different types of risks.

Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently. For example: "internet" and "users" will appear more often in documents produced by firms in the technology sector; "oil", "natural gas" and "drilling" will appear more frequently in documents produced by firms in the oil industry, while "company" and "cash" would appear similarly in both.

A document typically concerns multiple topics, or in this case risks, in different proportions; thus, in a company risk disclosure that is concerned with 20% about financial risks and 20% about internet operations, the risk report would approximately have around 8 times more technology words than financial words.

Because of the large number of firms in the stock market, the amount of time to read, categorize and quantify the risks disclosed by every firm is simply beyond human capacity, but topic models are capable of identifying these risks.

The most common topic model currently in use is the LDA model proposed by Blei, Ng, and Jordan (2003). The model generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers per-document discrete distributions over topics. The interaction between the observed documents and the hidden topic structure is manifested in the probabilistic generative process associated with LDA.

4.5 LDA

In LDA each document can be described by a (probability) distribution over topics and each topic can be described by a (probability) distribution over words. In matrix algebra terms, we are factorizing the term-document matrix D into a matrix W mapping words to topics, and a matrix T mapping topics to words, similar to the factorization used in Principal Component Analysis, see Figure 6. In this way, LDA reduces the dimensionality of each document, from thousands of words, to the number of topics (25 in our case). However, LDA retains most of the information about the individual word counts, since the topics themselves are probability

distribution over words

Formally, LDA is a Bayesian factor model for discrete data that considers a fixed latent set of topics. Suppose there are D documents that comprise a corpus of texts with V unique terms. The K topics (in this case, risk types), are probability vectors $\beta_k \in \Delta_{V-1}$ over the V unique terms in the data, where Δ_M refers to the M-dimensional simplex. By using probability distributions, we allow the same term to appear in different topics with potentially different weights. We can think of a topic as a weighted word vector that puts higher mass in words that all express the same underlying theme.[14]

In LDA, each document is described by a distribution of topics that appear in the document, so each document d has its own distribution over topics given by θ_d (in our case, how much each company discusses each type of risk). Within a given document, each word is influenced by two factors, the topics proportions for that document, θ_{dk}, and the probability measure over the words within the topics. Formally, the probability that a word in document d is equal to the nth term is $p_{dn}\theta_d^k$.

It is easier to frame LDA in the language of graphical models, see Figure 7. Where M is the set of all the documents; N is the number of words per document. Inside the rectangle N we see w: the words observed in document i, z: the random topic for the jth word for document i, θ: the topic distribution for document i. α: the prior distribution over topics intuitively controls the sparsity of topics within a document (i.e. how many topics we need to describe a document). β the prior distribution of words within a topic controls how sparse the topics are in terms of words (i.e. how many words we need to describe a topic). There is a trade-off between the sparsity of the topics, i.e. how specialize they are, and the number of topics.

14. See Blei, Ng, and Jordan (2003) and Hansen, McMahon, and Prat (2018)

4.5.1 Number of topics

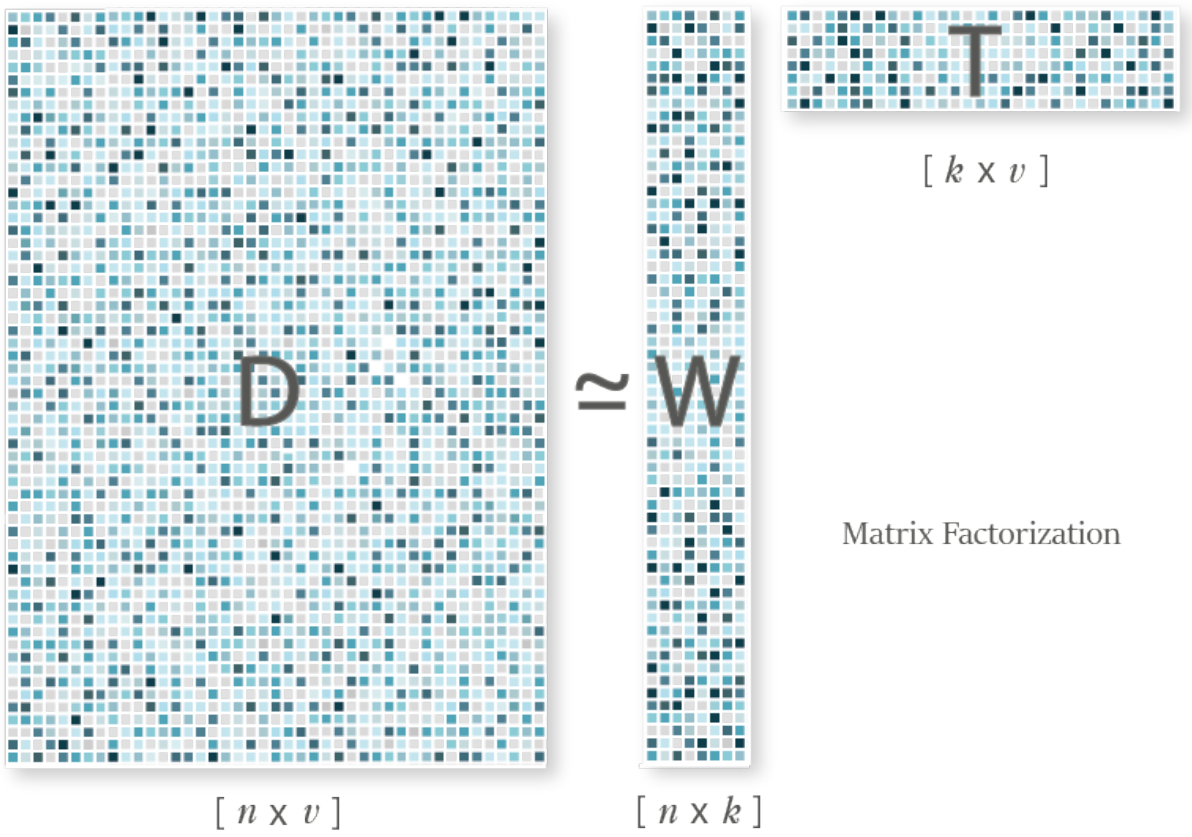
The number of topics is a hyperparameter in LDA. Ideally, there should be enough topics to be able to distinguish between themes in the text, but not so many that they lose their interpretability. I use the technical measure of topic coherence and out of sample log likelihood to help determine the optimal number of topics. In this case 25 topics accomplish this task, and is consistent with the numbers used in the literature of topic modeling in finance applications (Israelsen (2014), Bao and Datta (2014), Hanley and Hoberg (2019)).

A natural challenge is then to further reduce the extracted risks into a lower number of portfolios for the cross-section. See Section 8 for more details.

4.5.2 Estimation

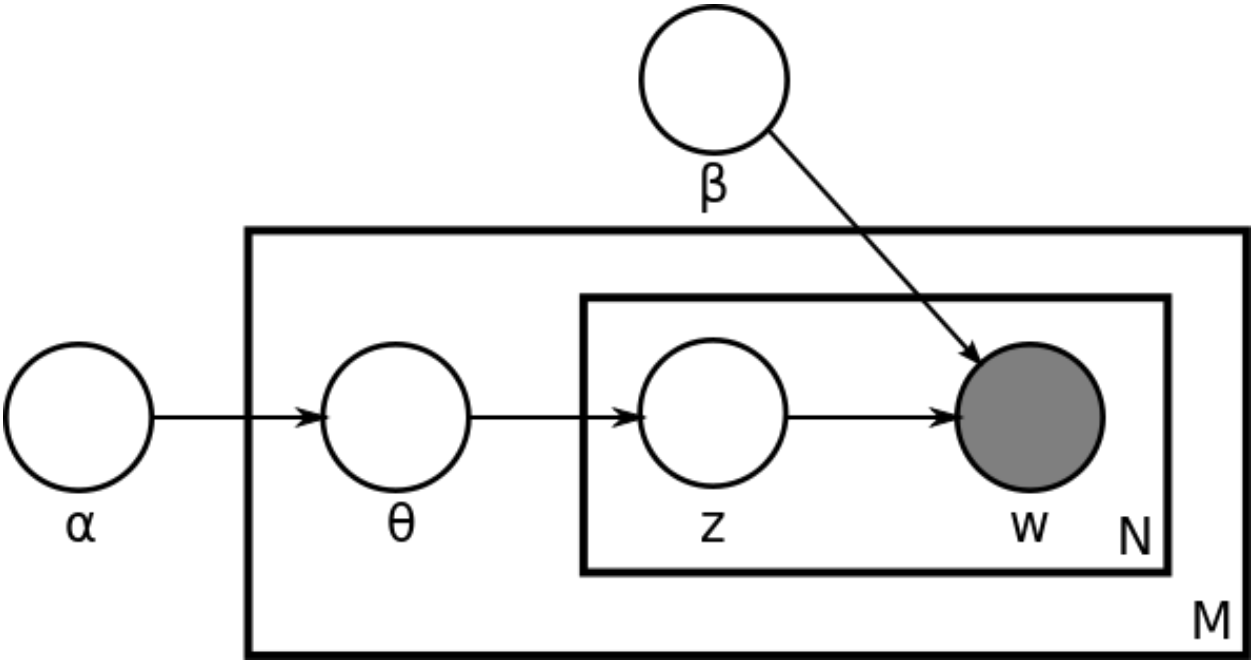
The estimation of the posterior parameters is done using the open-source software Gensim (Řehůřek and Sojka (2010)) which runs on Python. Gensim uses an online Variational Bayes algorithm. Because of the huge size of the collection of annual reports, the use of online algorithms allows us to not load every document into the RAM memory and hence we can estimate the model in a normal laptop. See the Appendix and Hoffman, Bach, and Blei (2010) for details. Because it is an online algorithm, the estimation is performed on a rolling basis. As new risk disclosures arrive, the risk topics get updated, and we get a new set of weights (the projection of the documents on the topic space).

Figure 6: Intuition for Topic Modelling



The figure shows the intuition for topic modeling. The Matrix D is the Document-Term Matrix with dimensions n x v, n is the number of documents and v is the number of terms. The matrix is intuitively decomposed into two matrices: Matrix T and Matrix W. Matrix T has dimensions k x v, where k is the number of topics and v is the number of terms. Each row in Matrix T sums up to one and all the elements are non-negative. Hence, each topic is a distribution over words. Matrix W has dimensions n x k, where k is the number of topics and n is the number of documents. Each row in Matrix W sums up to one and all the elements are non-negative. Hence, its rows are distributions over topics, risk weights.

Figure 7: LDA Graphical Model



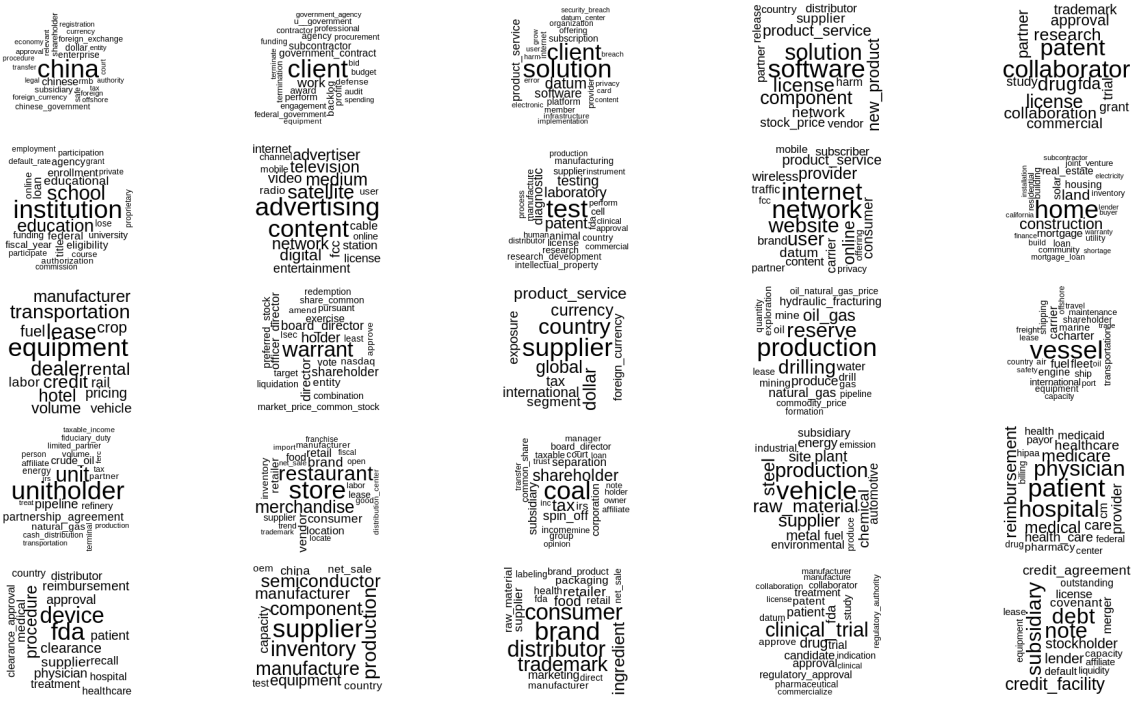
LDA in the language of graphical models. M is the set of all the documents; N is the number of words per document. Inside the rectangle N we see w: the words observed in document i, z: the random topic for the jth word for document i, θ: the topic distribution for document i. α: the prior distribution over topics intuitively controls the sparsity of topics within a document (i.e. how many topics we need to describe a document). β the prior distribution of words within a topic controls how sparse the topics are in terms of words (i.e. how many words we need to describe a topic).

5 Risk Topics

Recall that with LDA, we get in real-time all of the common risks that firms are talking about and how much time each company spends discussing each risk. Figure 10 shows an example of the latter.

To avoid confusion, I refer to the topics obtained using LDA as risk topics and to the amount of space they allocate to each risk as risk weights. Recall from Section 4 that risk topics are distribution over words, and risk weights are distribution over topics. It is important to remember that LDA is similar to a matrix factorization technique and does not give us labels for the risk topics, nevertheless, we can interpret the topics by reading the most frequent words as Figure 8 shows, and by looking at which companies discuss the most each risk.

Figure 8: Risk Topics

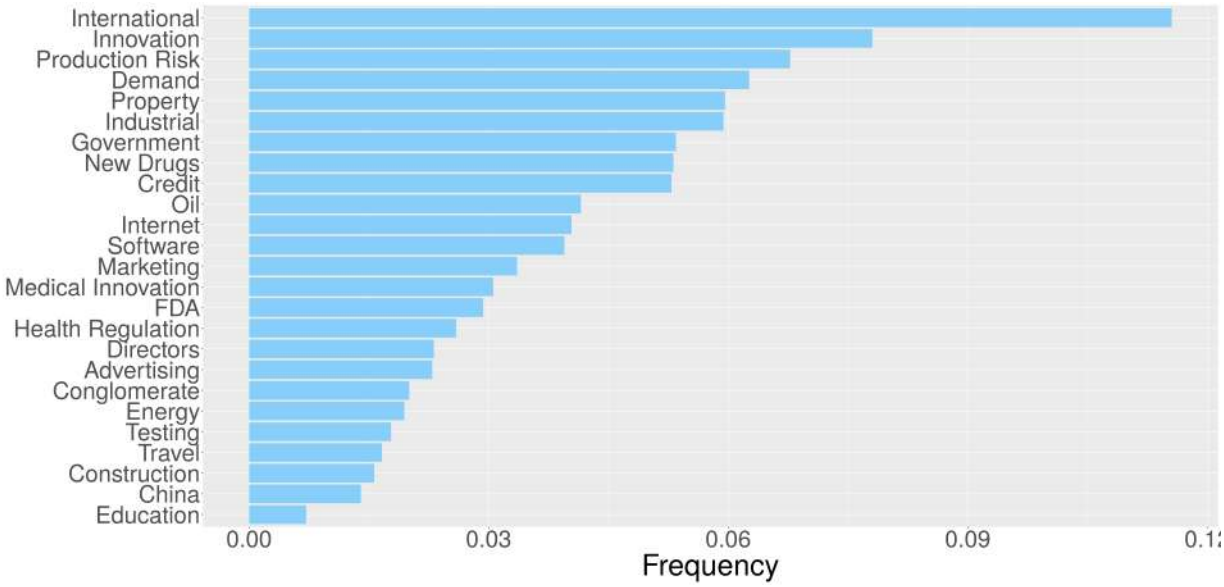


In this word cloud of the risks that firms face, a bigger font corresponds to a bigger weight for that word within each topic. See Section 4 for details on the procedure

Figure 8 shows a general picture of the risks that firms are concerned about, where I

show the 25 risk topics extracted from the 10-K annual reports.

Figure 9: Average percentage of the risk disclosure that firms allocates to each risk



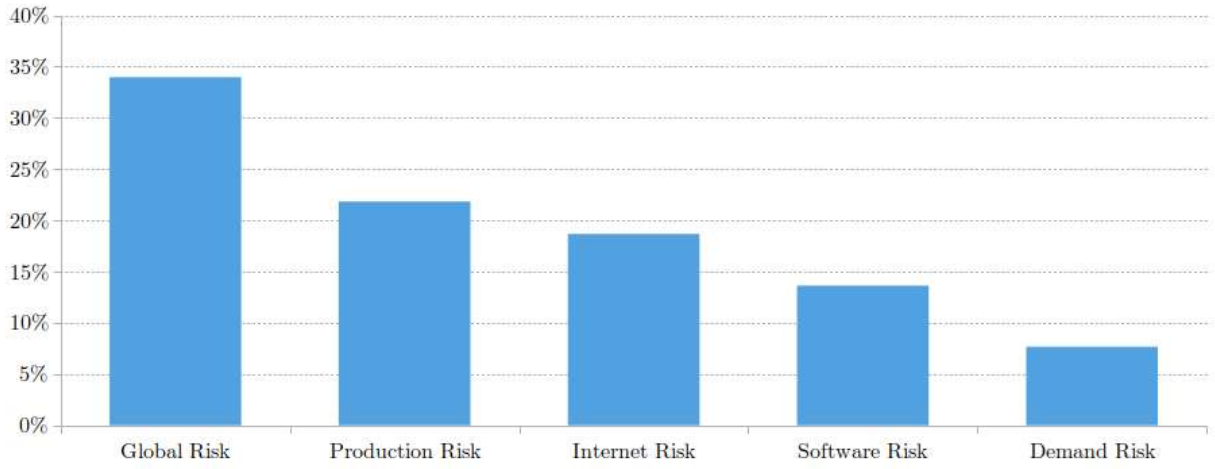
The figure shows the cross-sectional and time-series average of the percentage of the risk disclosure that firms allocates to each type of risk in the Section 1A: Risk Factors for the years 2006-2018. The values are obtained using Latent Dirichlet Allocation. See Sections 3 and 5 for details

Figure 11 shows that there is significant interaction between the risk weights and market beta, the book-to-market ratio and the market capitalization. There is no significant correlation between profitability, investment or past returns and any of the risks that firms disclose.

However, the factors in standard pricing models are portfolios of firms, and hence it is natural to wonder what is the implied risk disclosure for the portfolios. That is, if we think of a factor as a hypothetical firm, what would be its risk disclosure. Notice that to understand portfolios this way we need to have the composition of the portfolios. Once we construct factor-mimicking portfolios of each risk, we will be able to study the exposure for an arbitrary portfolio whose returns we have, using standard projection techniques.

Figure 12 shows the implied (time-series average) risk disclosure for the HML factor, constructed by using a weighted average of firms' risk disclosures where the weights are the portfolio weights. The implied HML factor risk disclosure is heavily short international risk,

Figure 10: Percentage of the risk disclosure that Apple Inc. allocates to each risk



The table shows the percentages of the risk disclosure that Apple Inc. allocates to each type of risk in the Section 1A: Risk Factors for their 2016 annual report. The table only shows the five most discussed risks. The values are obtained using Latent Dirichlet Allocation. See Sections 3 and 5 for details

marketing risk, and heavily long oil and property risks. The decline in oil prices explains a significant part of its poor performance during the recent period. Notice the figure does not show the betas with respect to the factor-mimicking portfolios, which look fairly similar, as we will see in Section 9.

Figure 13 shows the exercise repeated for the Momentum Factor (portfolios sorted on past performance). There is a clear pattern of no stable relationship with any of the firms' risks: Momentum is moving back and forth between all of the types of risks. We may initially think the effect is mechanical, since Momentum is re-balanced monthly, whereas the risk exposures are stable. However, Momentum could be concentrated in a specific risk, say, in international risk. We see it is not the case and, Momentum does not seem to be related to any of the companies' disclosed risks.

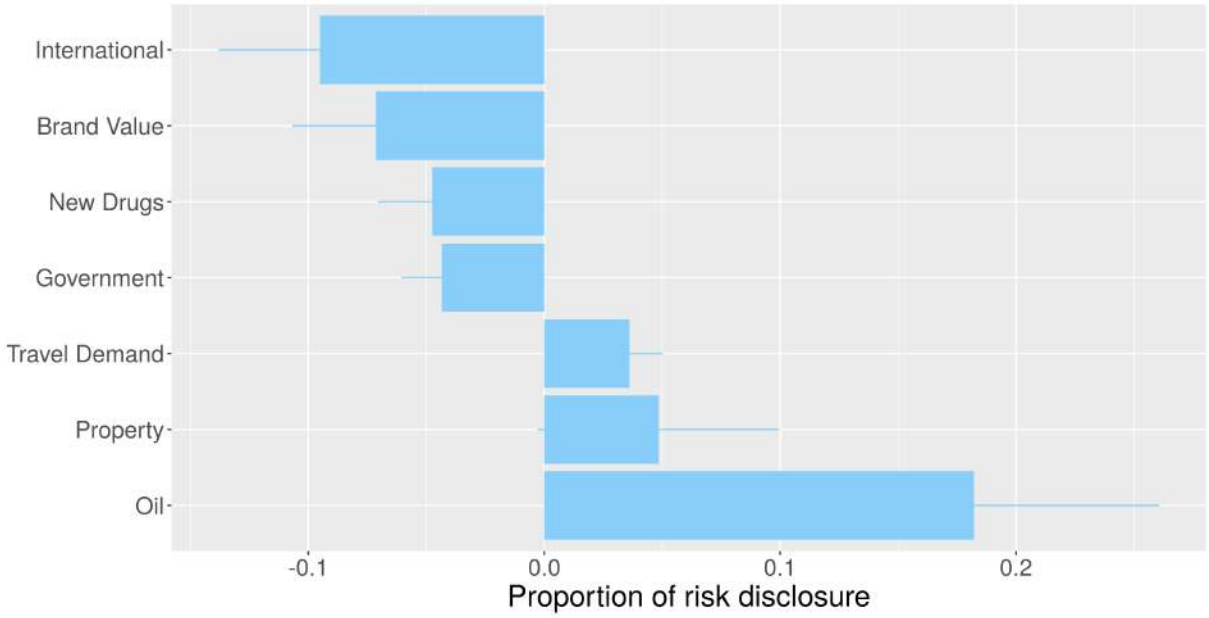
I describe here the four risks that affect the highest number of firms, as Figure 9 shows: Technology Risk, Production Risk, International Risk, and Demand Risk.[15] Firms allocate, on average, 36% of their risk disclosures discussing these four risks, and allocate the remaining

15. See the online Appendix for the rest.

Figure 11: Correlation of the risk weights with market beta, book-to-market, size, profitability, investment and past returns.

The Figure shows correlation between risk weights and some common predictors of the cross-section of returns. The sample period is 2006-2019. The predictors include yearly rolling window betas calculated with daily returns; and book-to-market ratios, size, profitability, and investment calculated as in Fama and French (2015). The data comes from the merged CRSP/Compustat database and the 10-K reports. The risk weights are calculated as in Section 4.

Figure 12: Implied average percentage of the risk disclosure that the High-minus-low (book-to-market) factor allocates to each risk



The figure shows the time-series average of the percentage of the weighted proportions of the risk disclosures that firms allocates to each type of risk in the Section 1A: Risk Factors for the years 2006-2018. The weights correspond to the weights in the High-minus-low (book-to-market) factor. The lines are one standard deviation from the average value. The values are obtained using Latent Dirichlet Allocation. See Sections 3 and 5 for details

64% to the other 21 risks. The firms that allocate more than 25% across the four risk topics, are about half of the firms in the sample as Table 1 shows.

5.1 Innovation and Innovation Risk

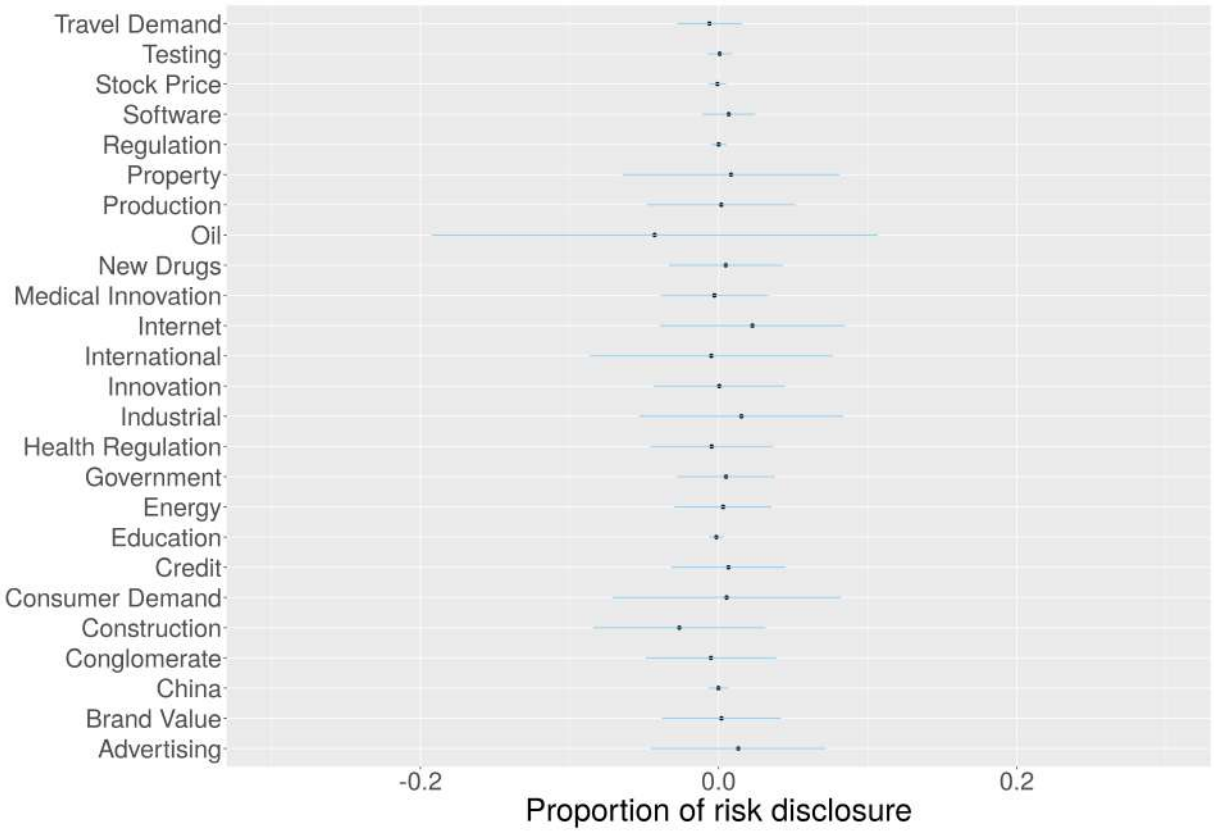
The most discussed risk topic, the Innovation Risk Topic, is characterized by words that have a direct relation to innovation, such as "software," "new product," "intellectual property,"

Table 1: Average proportion of the risk disclosures allocated to each risk for the most discussed risks in the year 2006

Innovation Risk	Production Risk	International Risk	Demand Risk	Total
0.11	0.09	0.08	0.08	0.36

The table shows the cross-sectional average of each firms distribution over topics for the annual reports of 2006, but only for the four most mentioned topics. See Section 4 for details.

Figure 13: Implied average percentage of the risk disclosure that the Momentum factor allocates to each risk



The figure shows the time-series average of the percentage of the weighted proportions of risk disclosures that firms allocates to each type of risk in the Section 1A: Risk Factors for the years 2006-2018. The weights correspond to the weights in the Momentum factor. The lines are one standard deviation from the average value. The values are obtained using Latent Dirichlet Allocation. See Sections 3 and 5 for details

and "network," as Figure 14 shows. Table 3 shows that when we inspect the largest companies that spend more than 25% of their risk disclosures commenting about the Innovation Risk Topic, we see companies that spend a lot of resources on innovation: Microsoft, Oracle, Cisco, HP, among others.

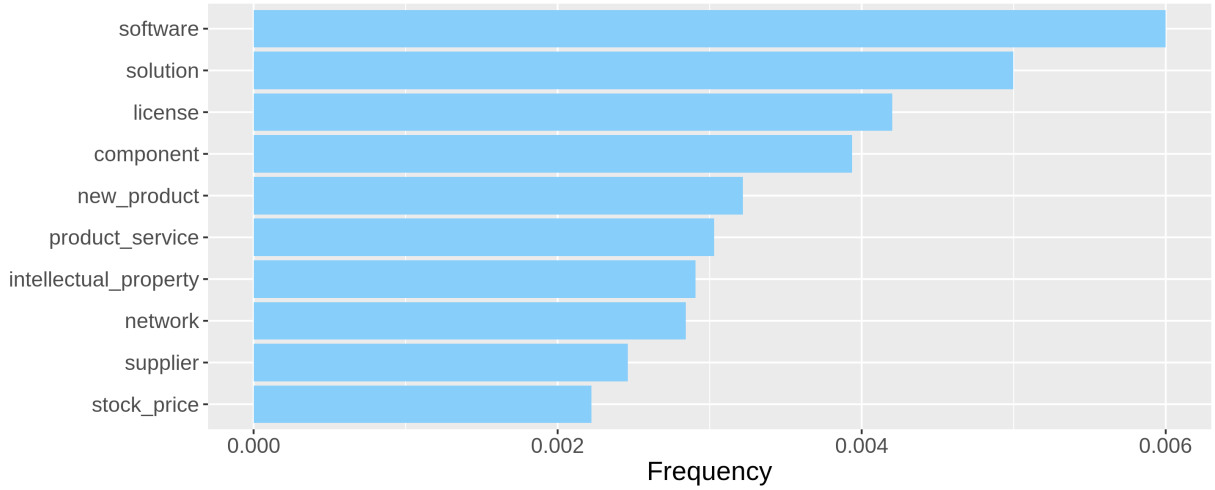
At this point, it is natural to wonder whether the risk topics are just capturing industry-specific risks. However, it is not the case. Table 4 shows that the SIC industries cannot fully capture the relationship between risks and industries. Although, as expected, the firms that load on the Innovation Risk are concentrated mostly in the electronic, computing, and business services sectors; the SIC codes are too rigid and put half of these firms in the

Table 2: Number of firms heavily exposed to each risk

Year	Innovation Risk	Production Risk	International Risk	Demand Risk	Percentage of Total Firms
2006	413	364	343	264	0.54
2007	442	354	324	245	0.52
2008	388	270	356	223	0.50
2009	355	305	412	215	0.51
2010	300	275	387	211	0.48
2011	285	266	422	221	0.49
2012	258	252	468	202	0.50
2013	261	237	452	205	0.49
2014	248	215	479	203	0.48
2015	230	197	493	196	0.46
2016	213	171	505	205	0.44

The table shows the number of firms that spend more than 25% of the time discussing each topic. See the text for details.

Figure 14: Innovation Risk Topic



Distribution of the 10 most frequent words for the Innovation and Innovation Risk Topic

Table 3: Largest 10 Companies that are exposed more than 25% to the Innovation and Innovation Risk Factor

Company Name	Market Value (Millions)
MICROSOFT CORP	354392
ORACLE CORP	166066
CISCO SYSTEMS INC	144516
QUALCOMM INC	81885
EMC CORP/MA	49896
HP INC	48628
ADOBE SYSTEMS INC	45530
ILLUMINA INC	28136
VMWARE INC -CL A	23870
ELECTRONIC ARTS INC	19873

The table shows the largest firms by market capitalization measured in June 2016 that allocate more than 25% of their risk disclosure to the Innovation Risk Topic. See the text for details.

manufacturing division and the other half in the services division. The main reason being that industry classification is about what the business of the firm is, so HP and Oracle will look very different in that perspective, one selling computers and the other selling software services. However, they share similar risks, mainly technical challenges, and innovation-related risks.

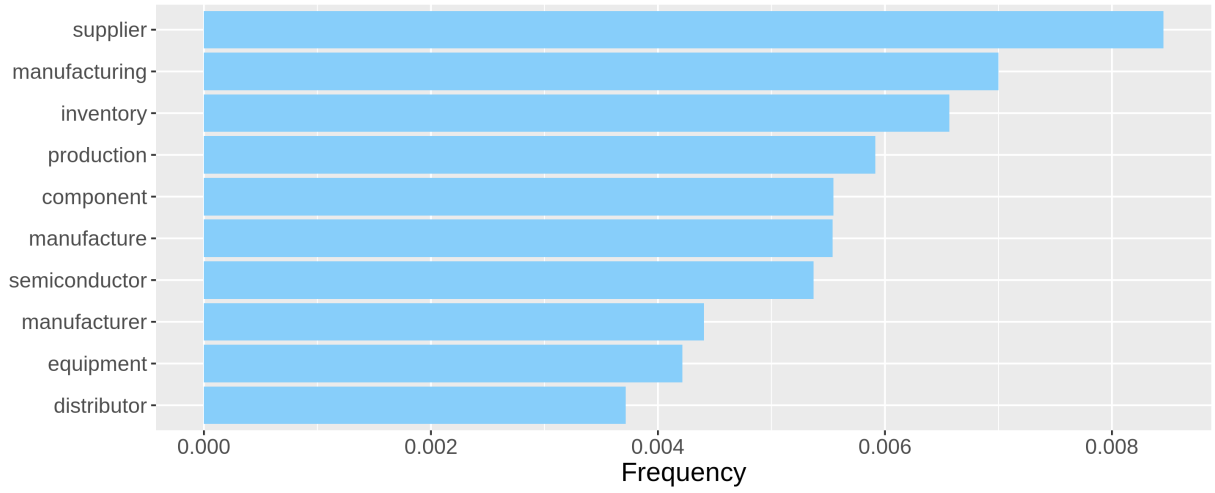
Notice we can think of the risks as spanning a more flexible Industry Classification where firms in comparable 'industries' face similar risks, instead of firms in related industries producing similar goods. Hoberg and Phillips (2016) pursue this idea and use a different section in the report, Item 1 Business Description, to create a more flexible Industry Classification. There does not exist a bijection from the risk topics to the text-based industries, as was the case with SIC industries. The difference comes mainly since two firms can be exposed to the same risk but be in completely unrelated industries, e.g., international risk, or be in the same industry but share different risks, e.g., two companies in the manufacturing sector, one in credit distress.

Table 4: Number of firms by SIC code for firms that are exposed to the Innovation Risk Factor

2-Digit SIC Code	Industry	Division	Number of firms
35	Manufacturing	Industrial and Commercial Machinery and Computer Equipment	43
36	Manufacturing	Electronic and other Electrical Equipment and Components, except Computer Equipment	58
38	Manufacturing	Measuring, Analyzing, and Controlling Instruments; Photographic, Medical and Optical Goods; Watches and Clocks	18
73	Services	Business Services	82

The table shows the number of firms by SIC code for the firms that allocate more than 25% of their risk disclosure to the Innovation Risk Topic. The number of firms is taken at June 2016. I only present the SIC codes for which the number of firms is higher than 15. See the text for details.

Figure 15: Production Risk Topic



Distribution of the 10 most frequent words for the Production Risk Topic

Table 5: Largest 10 Companies that are exposed more than 25% to the Production Risk

Company Name	Market Value (Millions)
INTEL CORP	162776
TEXAS INSTRUMENTS INC	55428
APPLIED MATERIALS INC	19453
ANALOG DEVICES	18761
WESTERN DIGITAL CORP	18037
UNDER ARMOUR INC	17419
MICRON TECHNOLOGY INC	17050
SKYWORKS SOLUTIONS INC	16025
NVIDIA CORP	15787
SEAGATE TECHNOLOGY PLC	14984

The table shows the largest firms by market capitalization measured in June 2016 that allocate more than 25% of their risk disclosure to the Production Risk Topic. See the text for details.

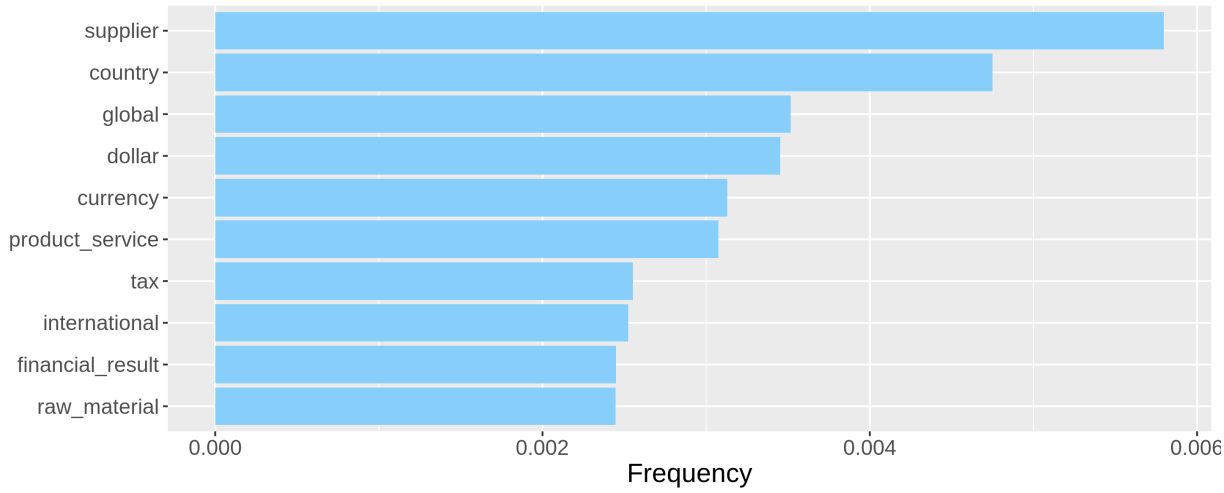
5.2 Production Risk

The Production Risk Topic is characterized by words that have a direct relation to production, such as "supplier", "manufacturing," "inventory," "component," as Figure 15 shows. Table 5 shows that when we inspect the largest companies that spend more than 25% of their risk disclosures commenting about the Production Risk Topic, we see companies whose production process seems to be very relevant for the business: Intel, Nvidia, Under Armour, among others.

5.3 International Risk

The International Risk Topic is characterized by words that have a direct relation to international concerns, such as "currency", "dollar," "global," "country," as Figure 16 shows. Table 6 shows that when we inspect the largest companies that spend more than 25% of their risk disclosures commenting about the International Risk Topic, we see companies that operate in global markets: Apple Inc, Exxon Mobile, Procter & Gamble, Coca-Cola, among others.

Figure 16: International Risk Topic



Distribution of the 10 most frequent words for the International Risk Topic (excluding "company")

5.4 Demand Risks

The Demand Risk Topic is characterized by words that are related to demand and sales, such as: "consumer," "store," "retail," and "merchandise," as Figure 17 shows. Table 7 shows that when we inspect the largest companies that spend more than 25% of their risk disclosures commenting about the Demand Risk Topic, we see the companies that focus on the consumption sector and hence are more prone to demand shocks: Wal-Mart, Home-Depot, McDonald's, Starbucks.

Table 6: Largest 10 Companies that are exposed more than 25% to the International Risk

Company Name	Market Value (Millions)
APPLE INC	615336
EXXON MOBIL CORP	323960
PROCTER & GAMBLE CO	212388
AT&T INC	211447
PFIZER INC	199329
COCA-COLA CO	185759
CHEVRON CORP	169378
ORACLE CORP	166066
INTEL CORP	162776
MERCK & CO	146899

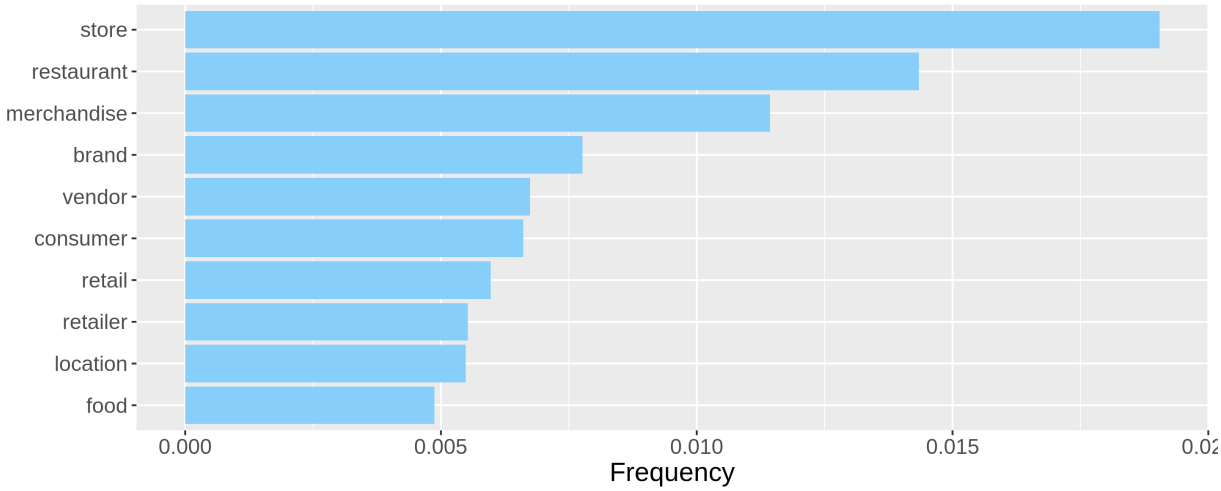
The table shows the largest firms by market capitalization measured in June 2016 that allocate more than 25% of their risk disclosure to the International Risk Topic. See the text for details.

Table 7: Largest 10 Companies that are exposed more than 25% to Demand Risks

Company Name	Market Value (Millions)
WALMART INC	209830
HOME DEPOT INC	157452
MCDONALD'S CORP	107129
NIKE INC	92880
STARBUCKS CORP	84413
LOWE'S COMPANIES INC	65211
COSTCO WHOLESALE CORP	61335
TJX COMPANIES INC	47267
TARGET CORP	43613
KROGER CO	37529

The table shows the largest firms by market capitalization measured in June 2016 that allocate more than 25% of their risk disclosure to the Demand Risk Topic. See the text for details.

Figure 17: Demand Risk Topic



Distribution of the 10 most frequent words for the Demand Risk Topic

6 Systematic and Diversifiable Risks

Which risks are systematic? A first approach would be to consider the most discussed risks as systematic. However, just because many companies are talking about a particular risk, say, a possible shortage in the supply of computer parts, it does not mean it cannot be diversified away by investors. Nevertheless, an international war can trigger a shortage in the supply of components, which seems systematic.[16]

I design an econometric test to classify risks into systematic and idiosyncratic. Consider two firms: If each one of them is more exposed to a systematic risk factor in the Arbitrage Pricing Theory (APT) sense, the covariance between their returns will increase.[17]

Consider the model

$$r^e_{i,t+1} = \beta^{S\prime}_{i,t} f^S_{t+1} + \beta^{g\prime}_{i,t} g_{i,t+1} + \epsilon_{i,t+1}, \tag{1}$$

where f^s are systematic factors, g^i idiosyncratic components. The idiosyncratic components are not priced, but are reported by the firms.

What is θ_j, the risk topic proportion capturing? Section 3 presents evidence that θ is the relative risk exposure for a firm, including both systematic and idiosyncratic components. Hence

$$\theta_j \approx \frac{\beta_j}{\beta^{S\prime}_i \mathbf{1}_S + \beta^{g\prime}_i \mathbf{1}_g}, \tag{2}$$

where $\mathbf{1}_k$ is a vector of ones of size k.

Let $B_i = \beta^{S\prime}_i \mathbf{1}_S + \beta^{g\prime}_i \mathbf{1}_g$, and hence $\beta_i = \theta_i B_i$. Notice that because of the machine learning algorithm we are capturing relative risk exposures only if the exposure is positive, and we only consider companies with positive relative risk exposures and hence $B_i > 0$. We can

16. Furthermore, idiosyncratic risks can matter for the expected returns, not because of the loading in the stochastic discount factor which is naturally zero, but because of the effect on the timing and duration of the cash flows (see Grotteria (2019)) and the fact that the term structure of equity is rarely flat (see Binsbergen and Koijen (2017), Weber (2018), Bansal et al. (2019).

17. See Ross (1976). Notice also that the the R^2 from a time-series regression of returns on the proposed risk factor will be high.

write the covariance between returns as

$$cov(r_i, r_j) = \sum_{s'}^{S} \sum_{s}^{S} \beta_{s'}^{i} \beta_{s}^{j} \sigma_{k',k}, \tag{3}$$

where we have S number of systematic factors.

We have then that

$$\frac{\partial cov(r_i, r_j)}{\partial \beta_j^k \beta_i^k} = var(f_k) > 0 \tag{4}$$

if factor k is a systematic risk and

$$\frac{\partial cov(r_i, r_j)}{\partial \beta_j^g \beta_i^g} = 0 \tag{5}$$

if 'factor' g, is in fact, not a factor.

Furthermore ,

$$\frac{\partial cov(r_i, r_j)}{\partial \theta_j^k \theta_i^k} = B_i B_j var(f_k) > 0 \tag{6}$$

if factor k is a systematic risk and

$$\frac{\partial cov(r_i, r_j)}{\partial \theta_j^k \theta_i^k} = 0 \tag{7}$$

if 'factor' g, is in fact, not a factor.

Hence we can run regressions of the form

$$cov(r_i, r_j)_{t+1} = a + b\theta_{j,t}^k \theta_{i,t}^k + controls_t \tag{8}$$

where $\theta_j^k \theta_i^k$ is the product of the weight in the k risk for companies j and i, and test whether the b coefficient is greater than zero. The covariance is also estimated, so we need to use either Generalized Method of Moments (GMM) to estimate jointly the regression coefficients and the covariances, or bootstrap the standard errors. I choose the later since there are literally millions of pairwise covariances. Notice also that the regression can be

run at every point in time, but I focus in the full-sample estimation for the remainder of the Section and use the time-series average risk weight for each firm.

Table 8 show the results of running those regressions. I only show the systematic risks, the rest are idiosyncratic. The controls include common industry membership, market betas, size distance and book-to-market ratios distance. The interpretation changes with the controls, since every factor model can be rotated. For example, when controlling for the market betas, it is equivalent (up to measurement error) to orthogonalizing the factors with respect to the market portfolio. Hence, the interpretation changes: a risk is systematic after removing the common variation of the market. For the period 2006-2019 exposure to China and Industrial risk is systematic across all specifications. Oil, Innovation and Credit is only systematic without removing the common variation with the market. International is systematic after taking into account the common variation with the market.

Table 8: Systematic Risks

Risks	Coefficient	t-stat	Coefficient	t-stat	Coefficient	t-stat
China	34.23	6.84	24.16	5.37	24.14	5.77
Industrial	24.97	14.95	7.65	4.92	5.98	4.13
Production	11.05	6.85	0.84	0.57	0.02	0.018
Oil	9.30	6.67	-3.46	-2.70	-2.72	-2.27
Innovation	3.53	3.53	-0.82	-0.61	-1.24	-1.00
Credit	5.02	1.91	3.22	1.36	3.62	1.65
Regulation	2.90	1.93	-8.14	-2.90	-3.38	-1.29
International	2.02	1.29	3.52	2.51	1.03	1.02
Controls	No		Beta		All	

The table shows the result of testing which risks are systematic. The test consist of running regressions of the form $cov(r_i, r_j) = a + b\theta_j^k \theta_i^k + controls$ and verifying that the coefficient b is positive. A higher coefficient indicates a higher impact on the covariance of returns. The standard errors are corrected for the fact that the covariance is estimated using bootstrap. The controls include the product of betas where betas are calculated with yearly rolling window regressions with respect to the market using daily returns; book-to-market ratios and size distances calculated as in Fama and French (2015); and common industry membership defined as the two stocks being in the same 3-digit SIC code.

7 Price of Risk

Section 6 shows which risks are systematic and which risks are idiosyncratic. However, a risk may be systematic and have zero price of risk, or there may be a characteristic that does not generate significant correlation between returns, but is related to the cross-section of returns because it covaries with a priced source of risk (e.g. as in ICAPM and not in APT).[18] The natural way to look at the effect is with Fama-MacBeth regressions. The controls include betas, book-to-market ratios, size, profitability, and investment. Notice that since the risk weights sum up to one for each company, we cannot include an intercept in the regression since that would induce collinearity. [19]

I include in the regressions only stocks above the 20th percentile of size of the NYSE exchange and exclude microcaps to alleviate any concern about liquidity, see Fama and French (2008).[20]

Caution is required in interpreting the magnitude of the coefficients. Each coefficient shows the average excess return of a portfolio whose average risk disclosure is concentrated in each risk. However, the average standard deviation for a given risk weight is about 10%, so all of the coefficients should be divided by ten to get a rough sense of the marginal effect of the risk weights in the expected returns.

Innovation related risks carry the highest significant unconditional premium for this period. Since these firms have a low book-to-market the time series regressions when using the Fama-French Five-Factor model as controls will show a significant intercept. Nevertheless, the Sharpe ratio will not be excessively high, since the returns comes with an increase in covariance as Section 6 shows.

18. Merton (1973)

19. Alternatively, we can drop one of the risk weights, but then the interpretation changes significantly. Another alternative involves scaling the risk weights by the (estimated in the period before) standard deviation of each stock, under strong conditions of the covariance structure, this can approximate the actual beta.

20. Results are similar, although the coefficients are naturally slightly smaller when using only the big stocks, stocks above the 50th percentile of size of the NYSE exchange. Results are available in the online Appendix.

Table 9: Fama-MacBeth Regressions

	Excess Returns	
	(1)	(2)
China	1.085	1.042
	(0.961)	(1.062)
Government	0.792	1.122
	(1.831*)	(1.967**)
Software	1.765	2.137
	(3.608***)	(3.416***)
Innovation	1.275	1.520
	(2.520**)	(2.723***)
Controls	*No*	*Yes*

The table shows the monthly time-series average of the coefficients, i.e. the Fama-MacBeth estimates, for the risk weights we get from running cross-sectional regressions of the form $r^e_{i,t+1} = b'_t\theta_{i,t} + \gamma'_t x_{i,t}$ every month, where $r^e_{i,t+1}$ are the excess returns next period, $\theta_{i,t}$ are the risk weights (the relative exposure to each risk), and $x_{i,t}$ are additional controls available at time t. The t-stats are shown in parenthesis underneath each coefficient. The sample period is 2006-2019. The controls include yearly rolling window betas calculated with daily returns; and book-to-market ratios, size, profitability, and investment calculated as in Fama and French (2015). The data comes from the merged CRSP/Compustat database and the 10-K reports. The risk weights are calculated as in Section 4.

Consumer demand and international risk carry a premium when there are additional controls in the regressions. Production and Credit risk are marginally significant. The exposure to China and Oil Risk is not compensated in this period.

Overall, Innovation, Credit and International Risk, carry both a high risk premium and a high covariance, across specifications, despite the addition of controls, suggesting that they provide additional information both for first and second moments of returns.

Table 9: Fama-MacBeth Regressions (Continued)

	Excess Returns	
	(1)	(2)
Regulation	1.486	1.564
	(1.608)	(1.666*)
Education	0.540	1.131
	(0.820)	(1.419)
Advertising	0.621	1.021
	(1.094)	(1.450)
Testing	1.810	2.364
	(2.964***)	(3.332***)
Internet	1.293	1.718
	(2.889***)	(2.961***)
Construction	0.277	0.381
	(0.361)	(0.465)
Property	1.233	1.576
	(2.658***)	(2.491**)
Stock Price	−0.719	−0.698
	(−0.766)	(−0.799)
International	0.796	1.359
	(1.756*)	(2.138**)
Oil	0.343	0.467
	(0.393)	(0.542)
Controls	*No*	*Yes*

The table shows the monthly time-series average of the coefficients, i.e. the Fama-MacBeth estimates, for the risk weights we get from running cross-sectional regressions of the form $r^e_{i,t+1} = b'_t\theta_{i,t} + \gamma'_t x_{i,t}$ every month, where $r^e_{i,t+1}$ are the excess returns next period, $\theta_{i,t}$ are the risk weights (the relative exposure to each risk), and $x_{i,t}$ are additional controls available at time t. The t-stats are shown in parenthesis underneath each coefficient. The sample period is 2006-2019. The controls include yearly rolling window betas calculated with daily returns; and book-to-market ratios, size, profitability, and investment calculated as in Fama and French (2015). The data comes from the merged CRSP/Compustat database and the 10-K reports. The risk weights are calculated as in Section 4.

Table 9: Fama-MacBeth Regressions (Continued)

	Excess Returns	
	(1)	(2)
Travel Demand	0.570	0.773
	(0.851)	(1.132)
Energy	0.928	1.356
	(1.799*)	(2.218**)
Consumer Demand	0.795	1.142
	(1.456)	(1.646*)
Conglomerate	1.019	1.499
	(1.876*)	(2.381**)
Industrial	1.249	1.138
	(1.649*)	(1.751*)
Health Regulation	0.842	1.301
	(1.898*)	(2.124**)
Controls	*No*	*Yes*

The table shows the monthly time-series average of the coefficients, i.e. the Fama-MacBeth estimates, for the risk weights we get from running cross-sectional regressions of the form $r^e_{i,t+1} = b'_t\theta_{i,t} + \gamma'_t x_{i,t}$ every month, where $r^e_{i,t+1}$ are the excess returns next period, $\theta_{i,t}$ are the risk weights (the relative exposure to each risk), and $x_{i,t}$ are additional controls available at time t. The t-stats are shown in parenthesis underneath each coefficient. The sample period is 2006-2019. The controls include yearly rolling window betas calculated with daily returns; and book-to-market ratios, size, profitability, and investment calculated as in Fama and French (2015). The data comes from the merged CRSP/Compustat database and the 10-K reports. The risk weights are calculated as in Section 4.

Table 9: Fama-MacBeth Regressions (Continued)

	Excess Returns	
	(1)	(2)
Medical Innovation	1.225	1.687
	(2.656***)	(2.636***)
Production	1.141	1.296
	(1.734*)	(1.951*)
Brand Value	1.187	1.916
	(3.125***)	(2.896***)
New Drugs	2.025	2.524
	(2.922***)	(3.210***)
Credit	1.228	1.137
	(1.689*)	(1.706*)
Controls	No	Yes

The table shows the monthly time-series average of the coefficients, i.e. the Fama-MacBeth estimates, for the risk weights we get from running cross-sectional regressions of the form $r^e_{i,t+1} = b'_t\theta_{i,t} + \gamma'_t x_{i,t}$ every month, where $^e_{i,t+1}$ are the excess returns next period, $\theta_{i,t}$ are the risk weights (the relative exposure to each risk), and $x_{i,t}$ are additional controls available at time t. The t-stats are shown in parenthesis underneath each coefficient. The sample period is 2006-2019. The controls include yearly rolling window betas calculated with daily returns; and book-to-market ratios, size, profitability, and investment calculated as in Fama and French (2015). The data comes from the merged CRSP/Compustat database and the 10-K reports. The risk weights are calculated as in Section 4.

8 Portfolios

I use factor mimicking portfolios designed to have unit exposure in one risk and zero in the other risks. The relative risk exposure we get from the topic modeling algorithm is similar to an indicator variable, and hence there is no natural short side. An analogy are industries: there is no 'short' side of the coal industry. Similarly, the opposite of disclosing exposure to demand risk is not disclosing demand risk exposure. Because of the sparsity of the machine learning algorithm, the 'short' side would consist of a well-diversified portfolio extremely similar to the market portfolio.[21]

I use the cross-sectional technique constructed in Fama (1976) and recently described in Back, Kapadia, and Ostdiek (2015) to get the portfolio weights. As in Back, Kapadia, and Ostdiek (2015) and Fama and French (2008) I include only stocks above the 20th percentile of size of the NYSE exchange and exclude microcaps to alleviate any concern about liquidity.[22]

Formally, the weights for the portfolio of risk k solve the following problem at every point in time:

$$min_{w_k} \; w_k' w_k \; s.t. \; w_k' X = e_i,$$

where X is a $n \times K$ matrix, n is the number of stocks, K is the number of risks (25), whose columns are the risk weights, how much time each company spends discussing each risk, each row corresponds to a firm observation, at a given point in time and e_i denotes the i-basis vector of $\mathbb{R}^K$.

The solution is available in analytical form:

$$w_k = X'(X'X)^{-1}e_k,$$

21. As an alternative I consider an indicator variable which is one for the risk that the company discusses the most and value-weight the portfolios. Results are similar and available in the online Appendix.

22. An alternative involves using a value-weighted procedure, equivalent to a weighted Fama-MacBeth regression using as weights the inverse of the size of the firm. See Kirby (2019) to see why the procedure is not recommended. Results are available in the online Appendix.

or if we collect the portfolio weights for each risk as a column in a Matrix W,

$$W = X'(X'X)^{-1},$$

and notice that $W'X = I_k$ as desired.[23]

Notice that the realized excess returns of the portfolios are the (normalized) slope coefficients on Fama-MacBeth regressions of excess returns on the risk weights. Notice also the portfolios are excess return portfolios. The big advantage of Fama's insight is that we can include additional variables in the X matrix when running the Fama-MacBeth regressions. When we include (ex-ante) market beta for example, we will be forming (ex-ante) beta neutral portfolios. In fact, Back, Kapadia, and Ostdiek (2015) show that getting the portfolio weights using cross-sectional regressions at every time period, and then running time-series regressions is the natural way to correct for the errors-in-variables problem that arises since betas are estimated.

Hence, we get two set of factor mimicking portfolios. I call simply 'firm identified risk factors' the portfolios that have unit exposure in one risk and zero in the other risks. I call 'orthogonal factors', the portfolios which in addition to having unit exposure in one risk and zero in the other risk, are orthogonal to portfolios formed on ex-ante betas and characteristics.

8.1 Are the risks spanned?

It is natural to wonder about the relationship between the usual models, for example Fama and French (2015) and the firm identified risk factors. We can see in Table 11 that the traditional models do not span these set of factors.The natural interpretation is that these factors contain additional information, and in fact combining them with the traditional ones leads to greater improvement of the description of cross-section of returns, although at the

23. The problem can be succinctly written as $min_W \sum_i^K e_i' W'W e_i$ s.t. $W'X = I_K$.

cost of mixing economic risks with proxies for other factors that affect the stock returns (e.g. Profitability).

The firm identified risk factors do not span the traditional ones as Table 12 shows. Table 10 shows that the intercept of the Profitability factor is the only one with a positive significant 'alpha' with respect to the firms identified risk factors and a t-stat of 3.44. The Small Minus Big Factor has a negative 'alpha' with a t-stat of -2.04. with The GRS tests and the implied p-values suggest that the factors are not completely in the span of each others.

Table 10: Projection of the Five-Factor Model plus momentum on the FIRFs

	Intercept	t-stat	R^2
$R_m - R_f$	.08	0.84	0.93
SMB	-0.27	-2.04	0.60
HML	0.10	0.62	0.49
RMW	0.36	3.43	0.42
CMA	0.14	1.31	0.27
Mom	-0.21	-0.65	0.45

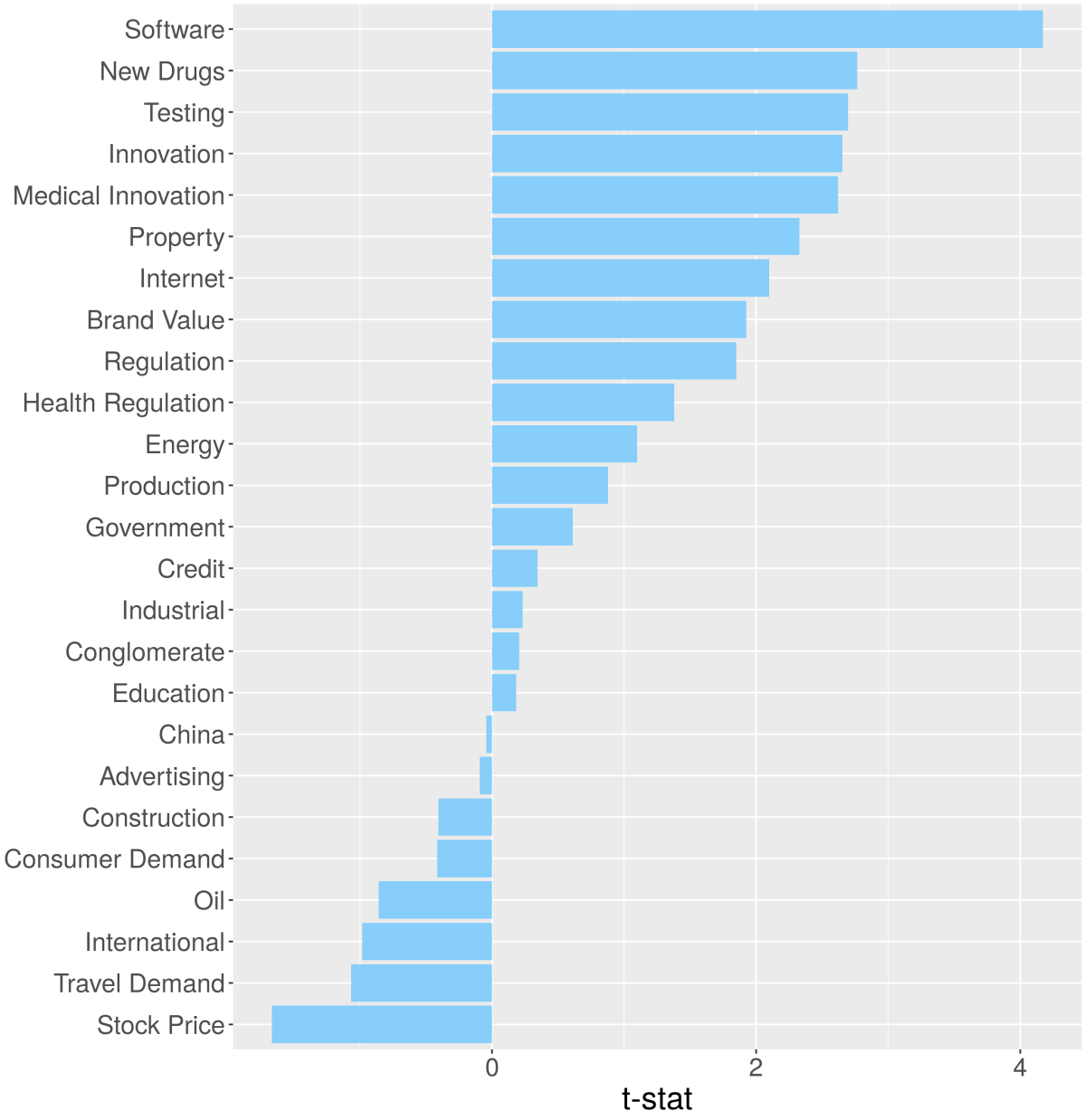
The Table shows the estimate of α_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the Fama-French Five-Factors, and the pricing factors f^e_{t+1} are the firm identified risk factors. $R_m - R_f$, the excess return on the market, SMB (Small Minus Big), HML (High Minus Low), RMW (Robust Minus Weak), CMA (Conservative Minus Aggressive), Mom (Momentum) are taken from French's website. R^2 is the adjusted coefficient of determination.

Figures 14-17 show the individual t-statistic confirming what we saw in the Fama-MacBeth Regressions: Risks related to innovation contain a significant component unexplained by the Fama-French Five-Factor Model. CAPM in this period performs significantly better than the Five-Factor Model, although there is still a significant unexplained component for the innovation risks.

Despite the models not being completely in the span of each others, it is interesting to see the betas between the firm identified risk factors and the standard factors. The firm identified risk factors are excess-return portfolios but not long-short, hence they are naturally correlated with the market portfolio, whereas the orthogonal factors have virtually zero exposure by design as Figure 22 shows.[24]

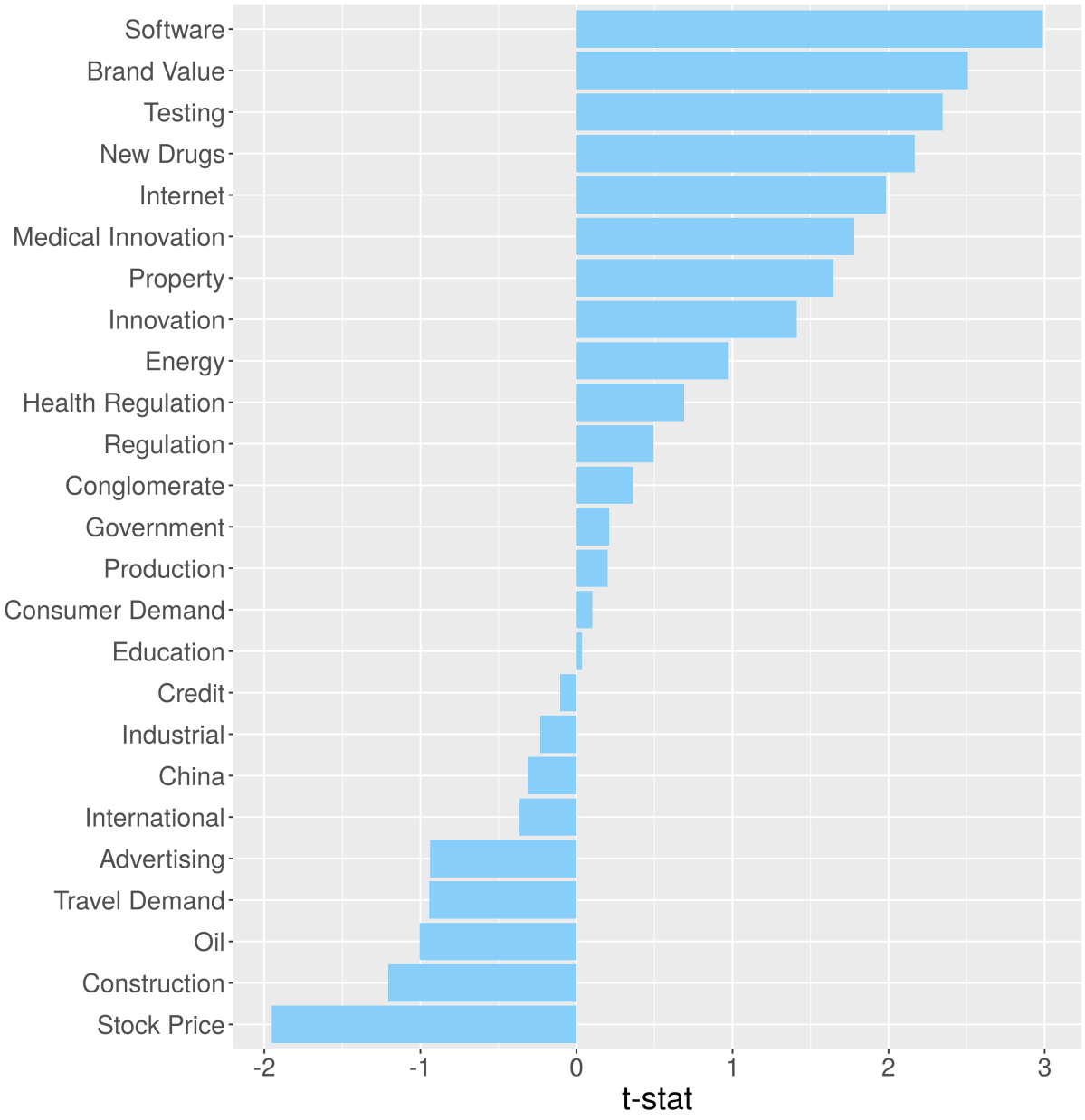
24. We can always rotate the FIRFs so that they are orthogonal to the market portfolio. The projec-

Figure 18: 'Alphas' of the firm identified risk factors with respect to the Fama-French Five-Factor Model (t-stats)



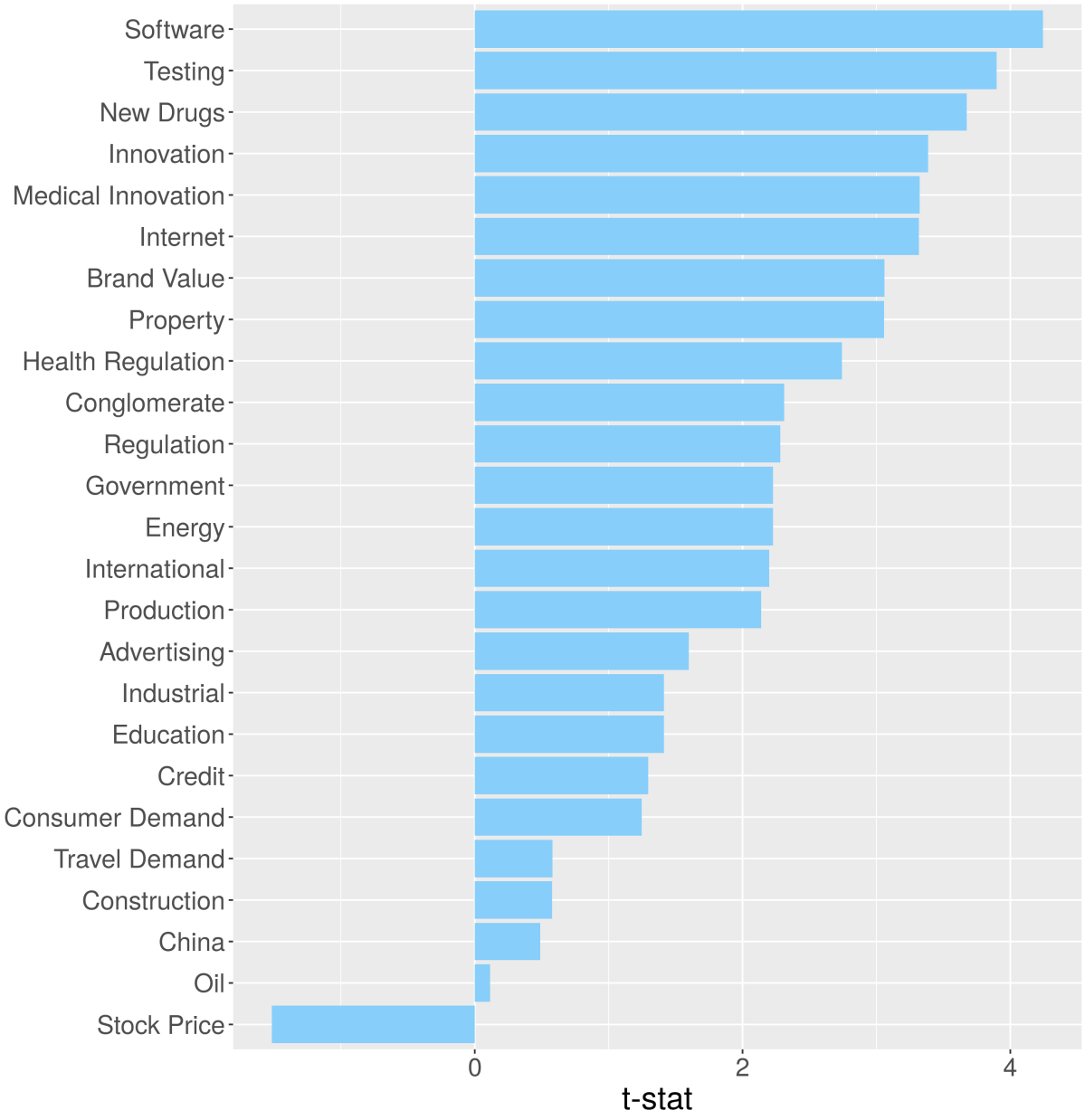
The Figure shows the t-stats of the coefficients α_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the firm identified risk factors, and the pricing factors f^e_{t+1} are the Fama-French Five-Factors. The standard errors are adjusted for heteroskedasticity and autocorrelation.

Figure 19: 'Alphas' of the firm identified risk factors with respect to CAPM (t-stats)



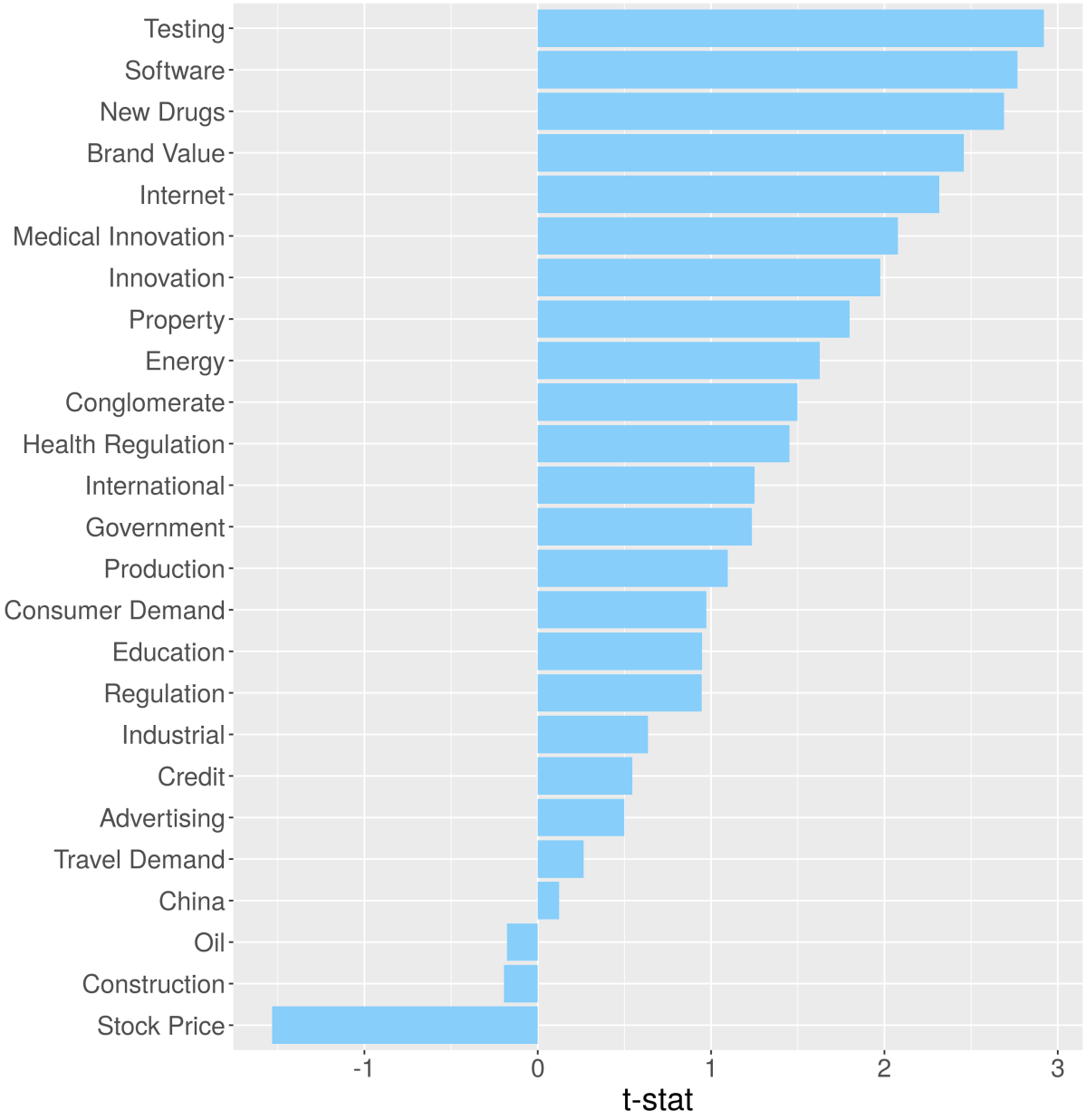
The Figure shows the t-stats of the coefficients α_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the firm identified risk factors, and the single pricing factor f^e_{t+1} is the excess return of the market portfolio. The standard errors are adjusted for heteroskedasticity and autocorrelation.

Figure 20: 'Alphas' of the orthogonal factors with respect to the Fama-French Five-Factor Model (t-stats)



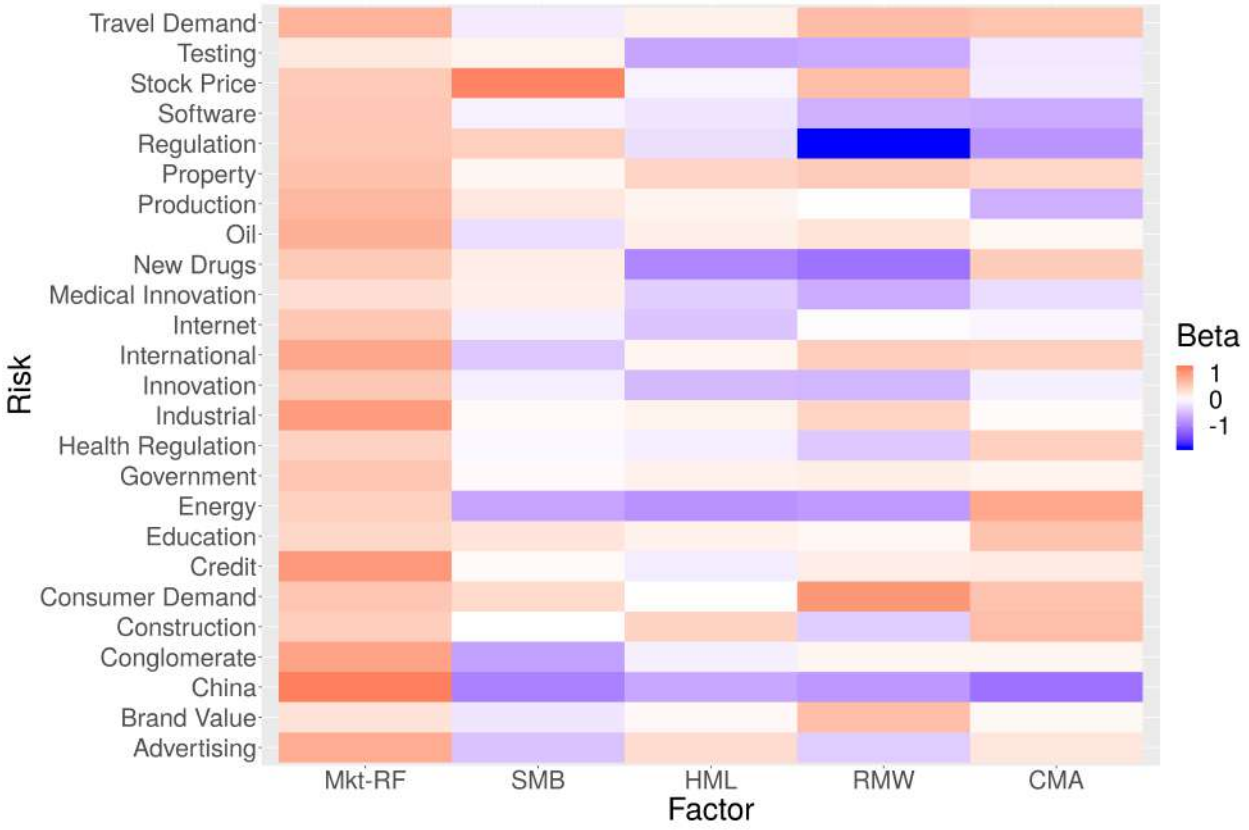
The Figure shows the t-stats of the coefficients α_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the orthogonal factors, and the pricing factors f^e_{t+1} are the Fama-French Five-Factors. The standard errors are adjusted for heteroskedasticity and autocorrelation.

Figure 21: 'Alphas' of the orthogonal factors with respect to CAPM (t-stats)



The Figure shows the t-stats of the coefficients α_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the orthogonal factors, and the single pricing factor f^e_{t+1} is the excess return of the market portfolio. The standard errors are adjusted for heteroskedasticity and autocorrelation.

Figure 22: Betas of the firm identified risk factors with respect to the Fama-French Five-Factor Model



The Figure shows the estimate of β_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the firm identified risk factors, and the pricing factors f^e_{t+1} are the Fama-French Five-Factors.

Table 11: GRS test: Are the FIRFs spanned?

	Firm Identified Risk Factors			Orthogonal Factors		
	GRS	p-value	R^2	GRS	p-value	R^2
CAPM	1.59	0.049	0.45	1.87	0.013	0.16
Fama-French 5 Factor Model	3.93	1.57e-7	0.57	1.98	0.007	0.49

The table shows the result of the GRS test: high values of the GRS statistic are indicative of high mispricing errors and generate a lower p-value, which is evidence against the fit of the model since the null is of no-mispricing. First row corresponds to using the market portfolio as the unique factor, second row corresponds to the Fama and French (2015) Factor Model. First column shows the result of the pricing of the firm identified risk factors. The second column shows the result of the pricing of the orthogonal factors. See the text for details.

Table 12: GRS test: Is the Five-Factor Model spanned?

	Fama-French Five-Factor Model		
	GRS	p-value	R^2
Firm Identified Risk Factors	6.83	2.6e-7	0.61

The table shows the result of the GRS test: high values of the GRS statistic are indicative of high mispricing errors and generate a lower p-value, which is evidence against the fit of the model since the null is of no-mispricing. The row corresponds to all Firm Identified Risk Factors as pricing factors.The column shows the result of the pricing of the Fama and French (2015) Factor Model by the firms identified risk factors.

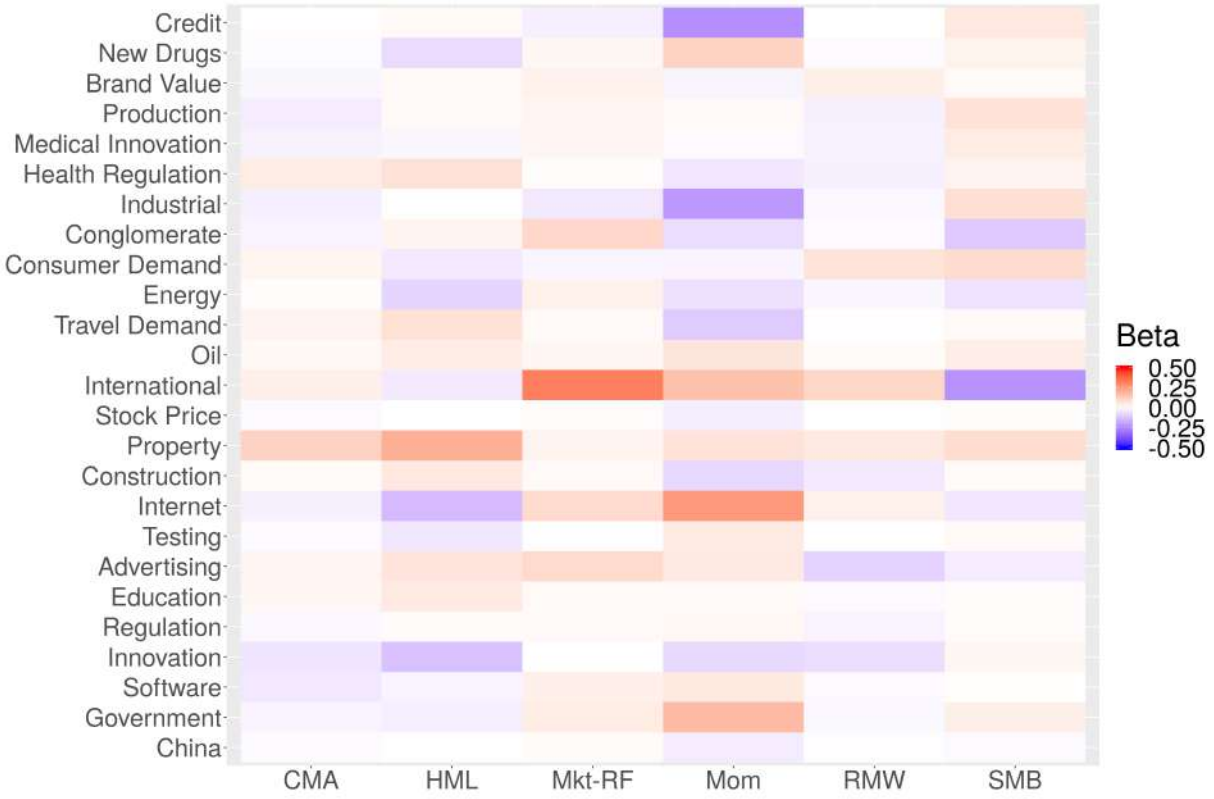
Figure 22 depicts the firm identified risk factors from the perspective of the Five-Factor Model. It shows that Stock-Price Risk covaries positively with the Small-minus-Big Factor, whereas International and China Risk covary negatively with it. Furthermore, the Innovation Risk Cluster covaries negatively with the High-minus-Low Factor and with the Profitability Factor, which explain the higher 'alphas' coming from the Five-Factor Model.

Figure 23 depicts the other side of the coin. The Five-Factor Model from the perspective of the firm identified risk factors. It shows that the market-portfolio consist mostly of International risk. Small Minus Big is negatively loaded on International Risk, and positively loaded in Consumer Demand and Production Risk.

High Minus Low is negatively loaded on Innovation and positively loaded on Oil, Production and Property Risk. RMW (Robust Minus Weak) is negatively related to Innovation

tions only makes sense for the the non-orthogonal factors, since by construction the orthogonal factors are orthogonal up to measurement error. The graphs are in Appendix 2 for completeness. See also Israelsen (2014).

Figure 23: Betas of the Fama-French Five-Factor Model (plus momentum) with respect to the firm identified risk factors



The Figure shows the estimate of β_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the Fama-French Five-Factors, and the pricing factors f^e_{t+1} are the firm identified risk factors.

Risk, and positively related with International and Demand Risk. CMA (Conservative Minus Aggressive) is positively related only to Property Risk. Momentum is negatively related to Credit and Industrial Risk and slightly positively related to the Innovation Risk Cluster.

Next, I discuss the how well the firm identified risk factors characterize the cross-section of stock returns.

9 Factor Performance

I select the 4 risks that affect the highest number of firms in 2006 and keep them for the whole sample to avoid look-ahead bias. Firms spend on average 36% of their risks disclosures discussing these 4 risks, and allocate the remaining 64% to the other 21 risks. Briefly, the risk factors correspond to Innovation Risk, Demand Risk, Production Risk and International Risk. I explore other dynamic approaches to select the factors in the Online Appendix. See Table 1.

Despite the model not being designed price the cross-section, it is interesting anyways to compare the performance of the factor model in pricing portfolios of general interest (such as the industry portfolios) and portfolios that are hard for macroeconomic based factors, for example the set of 25 book-to-market portfolios and the anomaly portfolios.[25] Adding more testing portfolios addresses the critique of Lewellen, Nagel, and Shanken (2010).

The table should not be read as a horse-race, other models are there just for comparability, since the models have different objectives, this one, to produce interpretable risk factors that represent economic risks for the firms. I use the GRS test from Gibbons, Ross, and Shanken (1989) and include the performance of the factor models of Fama and French (2015); Stambaugh and Yuan (2017); and Hou, Xue, and Zhang (2015) for benchmark comparison.

25. See Section 3 for details.

Recall that the GRS statistic is a measure of whether $\alpha_i = 0$ and that:

$$GRS \propto \frac{\alpha' \Sigma^{-1} \alpha}{1 + \mu' \Sigma^{-1} \mu}, \tag{9}$$

which we understand as a weighted and normalized sum of the squared alphas, divided by 1 plus the Sharpe ratio of the factors. Intuitively, if the test portfolios are spanned by the factors, we cannot increase the maximum Sharpe ratio that we get from the factors by adding the test portfolios and $\alpha_i = 0$.

High values of the GRS statistic are indicative of high mispricing errors ($|\alpha_i| \gg 0$), and low values are indicative of low mispricing errors ($\alpha_i \sim 0$). The null hypothesis in the GRS test is that the model is correct: there is no mispricing, the GRS statistic is small and $\alpha_i = 0$, hence, when the p-value is low we have strong evidence against the model and when the p-value is high, there is less evidence to reject the model. Lewellen, Nagel, and Shanken (2010) advice against the use of the average R^2 to make comparisons between factor models.

Table 13: GRS Test for the 4-Factor FIRFs Model and the Fama-French 5 Factor Model

	49 Industry + 25 B-to-M			49 Industry + 25 B-to-M + 15 α		
	GRS	p-value	R^2	GRS	p-value	R^2
4 FIRFs	1.53	0.03	0.69	2.043	0.007	0.639
Fama-French 5 Factor Model	1.69	0.01	0.76	2.271	0.003	0.731
Mispricing Factors	1.91	0.01	0.76	2.070	0.006	0.724
q-factor Model	1.62	0.02	0.73	2.328	0.002	0.704
All FIRFs	1.8	0.01	0.82	2.007	0.009	0.804
All FIRFs regularized	1.57	0.03	0.79	1.670	0.064	0.761

The table shows the result of the GRS test: high values of the GRS statistic are indicative of high mispricing errors and generate a lower p-value, which is evidence against the fit of the model since the null is of no-mispricing. Lewellen, Nagel, and Shanken (2010) advice against the use of the average R^2 to make comparisons between factor models. First row corresponds to the firm identified risk factors presented in the paper, second row corresponds to the Fama and French (2015) Factor Model, third row corresponds to the Anomaly Factors of Stambaugh and Yuan (2017), fourth row corresponds to the q-factor model of Hou, Xue, and Zhang (2015), fifth row corresponds to using all 25 of the risks, and sixth row corresponds to using all 25 of the risks and estimating the betas using LASSO regression. I perform the test on the joint set of 49 industry portfolios and 25 book-to-market portfolios available on Kennet French's website in the first column, and include the set of 11 long-short anomaly portfolios of Stambaugh and Yuan (2017) in the second column.

The firm identified risk factors is the best when we consider all portfolios jointly: the 49

industry portfolios, the 25 book-to-market portfolios, and the 11 anomaly portfolios. For the joint set of 25 book-to-market and 49 industry portfolios: The GRS statistic that measures whether $\alpha_i = 0$ is 1.52, lower than the GRS statistic of 1.85 for the Fama and French (2015) Model, and implies a p-value of 6.1%, so there is limited evidence against the model and $\alpha_i = 0$, hence, there is little evidence of mispricing; for comparison, the p-value for the Fama and French (2015) Model is 1.2%, that is, we can reject the null hypothesis that $\alpha_i = 0$ and there is evidence of mispricing. In short, the 4-factor model describes significantly better the joint set of 25 book-to-market and 49 industry portfolios than the leading factor models. The result is even sharper when we include the anomaly portfolios. See Table 13.

The model has an statistical fit significantly better than the factor models of Fama and French (2015); Stambaugh and Yuan (2017) and Hou, Xue, and Zhang (2015) in the set of 49 industry portfolios. Crucially, it explains the cross-sectional variation of returns: the GRS statistic that measures whether $\alpha_i = 0$ is .88, significantly lower than the GRS statistic of 1.55 for the Fama and French (2015) Model, and implies a p-value of 68%, that is, we cannot reject the null hypothesis that $\alpha_i = 0$, so there is little evidence of mispricing; for comparison, the p-value for the Fama and French (2015) Model is 4.4%, we can reject the null hypothesis that $\alpha_i = 0$ and there is stronger evidence of mispricing. In short, the GRS test says that the 4-factor model describes extremely well the set of expected returns of the 49 industry portfolios, especially compared to the factor models of Fama and French (2015), Stambaugh and Yuan (2017) and Hou, Xue, and Zhang (2015). See Table 14.

Surprisingly, the model has an statistical fit slightly better than the factor models of Fama and French (2015) and Hou, Xue, and Zhang (2015) in the test of the 25 book-to-market portfolios despite their inclusion of a book-to-market factor. The GRS statistic that measures whether $\alpha_i = 0$ is 1.83, slightly lower than the GRS statistic of 1.91 for the Fama and French (2015) Model. The factor model of Stambaugh and Yuan (2017) actually performs better, consistent with their evidence that book-to-market is not a proxy for risk, but rather for mispricing. Unfortunately, and as expected from the previous literature, there

is evidence of mispricing since the p-values are low for all of the models, recall that lower p-values imply there is more evidence against the models. See Table 14.

Table 14: GRS Test for the 4-Factor FIRFs Model and the Fama-French 5 Factor Model

	49 Industry Portfolios			25 Book-to-Market Portfolios			15 Anomaly Portfolios		
	GRS	p-value	R^2	GRS	p-value	R^2	GRS	p-value	R^2
FIRFs 4 Factor Model	0.88	0.679	0.63	1.83	0.019	0.8	1.34	0.21	0.21
Fama-French 5 Factor Model	1.55	0.045	0.68	1.91	0.013	0.94	1.12	0.35	0.43
Mispricing Factors	1.22	0.223	0.68	1.70	0.04	0.92	0.68	0.75	0.52
q-factor Model	1.47	0.073	0.67	1.88	0.02	0.92	1.13	0.35	0.43
All FIRFs	1.15	0.29	0.75	2.02	0.008	0.80	1.49	0.15	0.31
All FIRFs regularized	0.90	0.65	0.73	1.70	0.04	0.75	0.90	0.65	0.73

The table shows the result of the GRS test: high values of the GRS statistic are indicative of high mispricing errors and generate a lower p-value, which is evidence against the fit of the model since the null is of no-mispricing. Lewellen, Nagel, and Shanken (2010) advice against the use of the average R^2 to make comparisons between factor models. First row corresponds to the firm identified risk factors presented in the paper, second row corresponds to the Fama and French (2015) Factor Model, third row corresponds to the Anomaly Factors of Stambaugh and Yuan (2017), fourth row corresponds to the q-factor model of Hou, Xue, and Zhang (2015), fifth row corresponds to using all 25 of the risks, and sixth row corresponds to using all 25 of the risks and estimating the betas using LASSO regression. First and second columns correspond to the set of 49 industry portfolios and 25 book-to-market portfolios available on Kennet French's website, third column corresponds to the set of 11 long-short anomaly portfolios available of Stambaugh and Yuan (2017).

As an additional test I consider the anomaly portfolios from Stambaugh and Yuan (2017).[26] Naturally, the model of Stambaugh and Yuan (2017) performs best in these portfolios. A possible interpretation of the result is that most of these anomalies cannot be mapped to firms' risks and instead can be indicative of behavioral biases, market inefficiencies or be related to the SDF in dimensions other than risks that firms face. Table 14 shows that all the models are able to explain the cross-sectional differences in returns in the anomaly portfolios in the period 2006-2019 mainly because the performance of the anomalies has been declining, especially in the recent period as McLean and Pontiff (2016) document. The average R^2 is the lowest for the models using the firm identified risk factors, indicating that most of the anomalies do not covary significantly with any of the risks that firms are concerned about.

26. Available on their website.

10 Conclusion

I use machine learning to answer some of the essential questions in asset pricing: What are the fundamental risks in the economy? Which ones are systematic? Are they priced? Are they summarized well by existing models?

I identify the risks that firms consider relevant by letting firms themselves tell us what risks they face. I use natural language processing techniques to extract this information from their annual reports. Then, I design an econometric test to distinguish between systematic and idiosyncratic risks, and estimate that they contain information beyond the standard characteristics and factors. Furthermore, I introduce firm identified risk factors (FIRFs) that perform at least as well as traditional models while being literally described by words and not using any information from past prices.

I provide evidence that firms have a significant understanding of the risks they face, the information they provide is relevant to investors, and that it can provide guidance on how to improve our theoretical asset pricing models. Ultimately, this paper shows text analysis can help us understand investors' risk perception and their conditional information set.

References

Back, Kerry, Nishad Kapadia, and Barbara Ostdiek. 2015. "Testing Factor Models on Characteristic and Covariance Pure Plays." *Working Paper.* https://ssrn.com/abstract=2621696.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. "Measuring Economic Policy Uncertainty*." *The Quarterly Journal of Economics* 131, no. 4 (July): 1593–1636. ISSN: 0033-5533. doi:10.1093/qje/qjw024. eprint: http://oup.prod.sis.lan/qje/article-pdf/131/4/1593/30636768/qjw024.pdf. https://doi.org/10.1093/qje/qjw024.

Bansal, Ravi, Shane Miller, Dongho Song, and Amir Yaron. 2019. *The Term Structure of Equity Risk Premia.* Working Paper, Working Paper Series 25690. National Bureau of Economic Research, March. doi:10.3386/w25690. http://www.nber.org/papers/w25690.

Bao, Yang, and Anindya Datta. 2014. "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures." *Manage. Sci.* (Institute for Operations Research)(the Management Sciences (INFORMS), Linthicum, Maryland, USA) 60, no. 6 (June): 1371–1391. ISSN: 0025-1909. doi:10.1287/mnsc.2014.1930. https://doi.org/10.1287/mnsc.2014.1930.

Berk, Jonathan B., Richard C. Green, and Vasant Naik. 1999. "Optimal Investment, Growth Options, and Security Returns." *The Journal of Finance* 54 (5): 1553–1607. ISSN: 1540-6261. doi:10.1111/0022-1082.00161. http://dx.doi.org/10.1111/0022-1082.00161.

Binsbergen, Jules H. van, and Ralph S.J. Koijen. 2017. "The term structure of returns: Facts and theory." *Journal of Financial Economics* 124 (1): 1–21. ISSN: 0304-405X. doi:https://doi.org/10.1016/j.jfineco.2017.01.009. http://www.sciencedirect.com/science/article/pii/S0304405X17300223.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *J. Mach. Learn. Res.* 3 (March): 993–1022. ISSN: 1532-4435. http://dl.acm.org/citation.cfm?id=944919.944937.

Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu. 2019. "The Structure of Economic News." *Working Paper.* doi:http://dx.doi.org/10.2139/ssrn.3446225. https://ssrn.com/abstract=3446225.

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" [in English (US)]. *Review of Accounting Studies* 19, no. 1 (March): 396–455. ISSN: 1380-6653. doi:10.1007/s11142-013-9258-3.

Cochrane, John H. 1991. "Production-Based Asset Pricing and the Link Between Stock Returns and Economic Fluctuations." *The Journal of Finance* 46 (1): 209–237. ISSN: 1540-6261. doi:10.1111/j.1540-6261.1991.tb03750.x. http://dx.doi.org/10.1111/j.1540-6261.1991.tb03750.x.

———. 2005. *Asset Pricing: Revised Edition.* Princeton University Press. ISBN: 9781400829132. https://books.google.co.uk/books?id=20pmeMaKNwsC.

———. 2011. "Presidential Address: Discount Rates." *The Journal of Finance* 66 (4): 1047–1108. doi:10.1111/j.1540-6261.2011.01671.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2011.01671.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2011.01671.x.

Eisfeldt, Andrea L., and Dimitris Papanikolaou. 2013. "Organization Capital and the Cross-Section of Expected Returns." *The Journal of Finance* 68 (4): 1365–1406. ISSN: 1540-6261. doi:10.1111/jofi.12034. http://dx.doi.org/10.1111/jofi.12034.

Fama, Eugene F. 1976. *Foundations of Finance.* Basic Books.

Fama, Eugene F., and Kenneth R. French. 1992. "The Cross-Section of Expected Stock Returns." *The Journal of Finance* 47 (2): 427–465. ISSN: 1540-6261. doi:10.1111/j.1540-6261.1992.tb04398.x. http://dx.doi.org/10.1111/j.1540-6261.1992.tb04398.x.

———. 1993. "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics* 33 (1): 3–56. ISSN: 0304-405X. doi:https://doi.org/10.1016/0304-405X(93)90023-5. http://www.sciencedirect.com/science/article/pii/0304405X93900235.

———. 2008. "Average Returns, B/M, and Share Issues." *The Journal of Finance* 63 (6): 2971–2995. doi:10.1111/j.1540-6261.2008.01418.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2008.01418.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2008.01418.x.

———. 2015. "A five-factor asset pricing model." *Journal of Financial Economics* 116 (1): 1–22. ISSN: 0304-405X. doi:http://dx.doi.org/10.1016/j.jfineco.2014.10.010. http://www.sciencedirect.com/science/article/pii/S0304405X14002323.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu. 2017. "Taming the Factor Zoo."

Gaulin, Maclean. 2019. "Risk Fact or Fiction: The information content of risk factor disclosures." *Working Paper.*

Gibbons, Michael R., Stephen A. Ross, and Jay Shanken. 1989. "A Test of the Efficiency of a Given Portfolio." *Econometrica* 57 (5): 1121–1152. ISSN: 00129682, 14680262. http://www.jstor.org/stable/1913625.

Gomes, João, Leonid Kogan, and Lu Zhang. 2003. "Equilibrium Cross Section of Returns." *Journal of Political Economy* 111 (4): 693–732. doi:10.1086/375379. eprint: https://doi.org/10.1086/375379. https://doi.org/10.1086/375379.

Grotteria, Marco. 2019. "Follow the Money." *Working Paper*. doi:https://dx.doi.org/10.2139/ssrn.3281201. https://ssrn.com/abstract=3281201.

Hanley, Kathleen Weiss, and Gerard Hoberg. 2018. "Interpretation of Emerging Risks in the Financial Sector." *Forthcoming Review of Financial Studies*. https://ssrn.com/abstract=2792943.

———. 2019. "Dynamic Interpretation of Emerging Risks in the Financial Sector." *The Review of Financial Studies* (February). ISSN: 0893-9454. doi:10.1093/rfs/hhz023. eprint: http://oup.prod.sis.lan/rfs/advance-article-pdf/doi/10.1093/rfs/hhz023/28204249/hhz023.pdf. https://doi.org/10.1093/rfs/hhz023.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*." *The Quarterly Journal of Economics* 133 (2): 801–870. doi:10.1093/qje/qjx045. eprint: /oup/backfile/content_public/journal/qje/133/2/10.1093_qje_qjx045/1/qjx045.pdf. http://dx.doi.org/10.1093/qje/qjx045.

Harvey, Campbell R., Yan Liu, and Heqing Zhu. 2016. "... and the Cross-Section of Expected Returns." *The Review of Financial Studies* 29 (1): 5–68. doi:10.1093/rfs/hhv059. eprint: /oup/backfile/content_public/journal/rfs/29/1/10.1093_rfs_hhv059/2/hhv059.pdf. http://dx.doi.org/10.1093/rfs/hhv059.

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, no. 4 (August): 2135–2202. ISSN: 0033-5533. doi:10.1093/qje/qjz021. eprint: http://oup.prod.sis.lan/qje/article-pdf/134/4/2135/30044712/qjz021.pdf. https://doi.org/10.1093/qje/qjz021.

Hoberg, Gerard, and Gordon Phillips. 2016. "Text-Based Network Industries and Endogenous Product Differentiation." *Journal of Political Economy* 124 (5): 1423–1465. doi:10.1086/688176. eprint: https://doi.org/10.1086/688176. https://doi.org/10.1086/688176.

Hoffman, Matthew, Francis R. Bach, and David M. Blei. 2010. "Online Learning for Latent Dirichlet Allocation." In *Advances in Neural Information Processing Systems 23,* edited by J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, 856–864. Curran Associates, Inc. http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf.

Hou, Kewei, Chen Xue, and Lu Zhang. 2015. "Digesting Anomalies: An Investment Approach." *The Review of Financial Studies* 28 (3): 650–705. doi:10.1093/rfs/hhu068. eprint: /oup/backfile/content_public/journal/rfs/28/3/10.1093/rfs/hhu068/3/hhu068.pdf. +%20http://dx.doi.org/10.1093/rfs/hhu068.

Israelsen, Ryan D. 2014. "Tell It Like It Is: Disclosed Risks and Factor Portfolios." *Working paper.*

Jegadeesh, Narasimhan, and Di Wu. 2013. "Word power: A new approach for content analysis." *Journal of Financial Economics* 110 (3): 712–729. ISSN: 0304-405X. doi:https://doi.org/10.1016/j.jfineco.2013.08.018. http://www.sciencedirect.com/science/article/pii/S0304405X13002328.

Ke, Shikun, José Luis Montiel Olea, and James Nesbit. 2019. "A Robust Machine Learning Algorithm for Text Analysis." *Working Paper*.

Ke, Zheng, Bryan T. Kelly, and Dacheng Xiu. 2019. "Predicting Returns with Text Data." *University of Chicago, Becker Friedman Institute for Economics Working Paper*. doi:http://dx.doi.org/10.2139/ssrn.3074808. https://ssrn.com/abstract=3389884.

Kelly, Bryan, Seth Pruitt, and Yinan Su. 2018. "Characteristics Are Covariances: A Unified Model of Risk and Return," Working Paper Series, no. 24540 (April). doi:10.3386/w24540. http://www.nber.org/papers/w24540.

Kirby, Chris. 2019. "Firm Characteristics, Cross-Sectional Regression Estimates, and Asset Pricing Tests." Raz005, *The Review of Asset Pricing Studies* (June). ISSN: 2045-9920. doi:10.1093/rapstu/raz005. eprint: http://oup.prod.sis.lan/raps/advance-article-pdf/doi/10.1093/rapstu/raz005/29011928/raz005.pdf. https://doi.org/10.1093/rapstu/raz005.

Kogan, Leonid, and Dimitris Papanikolaou. 2014. "Growth Opportunities, Technology Shocks, and Asset Prices." *The Journal of Finance* 69 (2): 675–718. ISSN: 1540-6261. doi:10.1111/jofi.12136. http://dx.doi.org/10.1111/jofi.12136.

Kozak, SERHIY, STEFAN Nagel, and SHRIHARI Santosh. 2018. "Interpreting Factor Models." *The Journal of Finance* 73 (3): 1183–1223. doi:10.1111/jofi.12612. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.12612. https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12612.

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken. 2010. "A skeptical appraisal of asset pricing tests." *Journal of Financial Economics* 96 (2): 175–194. ISSN: 0304-405X. doi:https://doi.org/10.1016/j.jfineco.2009.09.001. http://www.sciencedirect.com/science/article/pii/S0304405X09001950.

Livdan, Dmitry, Horacio Sapriza, and Lu Zhang. 2009. "Financially Constrained Stock Returns." *The Journal of Finance* 64 (4): 1827–1862. ISSN: 1540-6261. doi:10.1111/j.1540-6261.2009.01481.x. http://dx.doi.org/10.1111/j.1540-6261.2009.01481.x.

Loper, Edward, and Steven Bird. 2002. "NLTK: The Natural Language Toolkit." In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1,* 63–70. ETMTNLP '02. Philadelphia, Pennsylvania: Association for Computational Linguistics. doi:10.3115/1118108.1118117. https://doi.org/10.3115/1118108.1118117.

Loughran, TIM, and BILL McDonald. 2016. "Textual Analysis in Accounting and Finance: A Survey." *Journal of Accounting Research* 54 (4): 1187–1230.

Loughran, Tim, Bill McDonald, and Ioannis Pragidis. 2019. "Assimilation of Oil News Into Prices." *Working Paper.* doi:http://dx.doi.org/10.2139/ssrn.3074808. https://ssrn.com/abstract=3074808.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* New York, NY, USA: Cambridge University Press. ISBN: 0521865719, 9780521865715.

McLean, R. David, and Jeffrey Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability?" *The Journal of Finance* 71 (1): 5–32. doi:10.1111/jofi.12365. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.12365. https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12365.

Merton, Robert C. 1973. "An Intertemporal Capital Asset Pricing Model." *Econometrica* 41 (5): 867–887. ISSN: 00129682, 14680262. http://www.jstor.org/stable/1913811.

Nagel, Stefan. 2005. "Short sales, institutional investors and the cross-section of stock returns." *Journal of Financial Economics* 78 (2): 277–309. ISSN: 0304-405X. doi:https://doi.org/10.1016/j.jfineco.2004.08.008. http://www.sciencedirect.com/science/article/pii/S0304405X05000735.

Řehůřek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora" [in English]. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* 45–50. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May.

Ross, Stephen A. 1976. "The arbitrage theory of capital asset pricing." *Journal of Economic Theory* 13 (3): 341–360. ISSN: 0022-0531. doi:https://doi.org/10.1016/0022-0531(76)90046-6. http://www.sciencedirect.com/science/article/pii/0022053176900466.

Stambaugh, Robert F., and Yu Yuan. 2017. "Mispricing Factors." *The Review of Financial Studies* 30 (4): 1270–1315. doi:10.1093/rfs/hhw107. eprint: /oup/backfile/content_public/journal/rfs/30/4/10.1093_rfs_hhw107/2/hhw107.pdf. +%20http://dx.doi.org/10.1093/rfs/hhw107.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–288. ISSN: 00359246. http://www.jstor.org/stable/2346178.

Weber, Michael. 2018. "Cash flow duration and the term structure of equity returns." *Journal of Financial Economics* 128 (3): 486–503. ISSN: 0304-405X. doi:https://doi.org/10.1016/j.jfineco.2018.03.003. http://www.sciencedirect.com/science/article/pii/S0304405X18300667.

Zhang, Lu. 2005. "The Value Premium." *The Journal of Finance* 60 (1): 67–103. ISSN: 1540-6261. doi:10.1111/j.1540-6261.2005.00725.x. http://dx.doi.org/10.1111/j.1540-6261.2005.00725.x.

Hanley , Kathleen Weiss and Hoberg, Gerard, Dynamic Interpretation of Emerging Risks in the Financial Sector (February 28, 2018). Available at SSRN: https://ssrn.com/abstract=2792943 or http://dx.doi.org/10.2139/ssrn.2792943

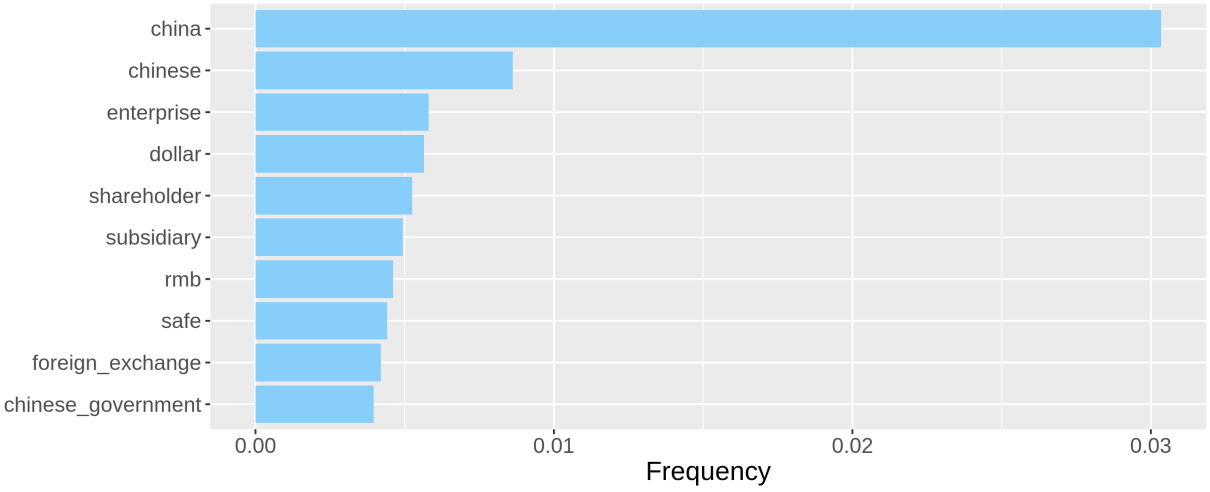
Hassan, Tarek A. and Hollander, Stephan and van Lent, Laurence and Tahoun, Ahmed, Firm-Level Political Risk: Measurement and Effects (December 2017). Available at SSRN: https://ssrn.com/abstract=2838644 or http://dx.doi.org/10.2139/ssrn.2838644

11 Appendix

Full online appendix available by request: joselop@wharton.upenn.edu

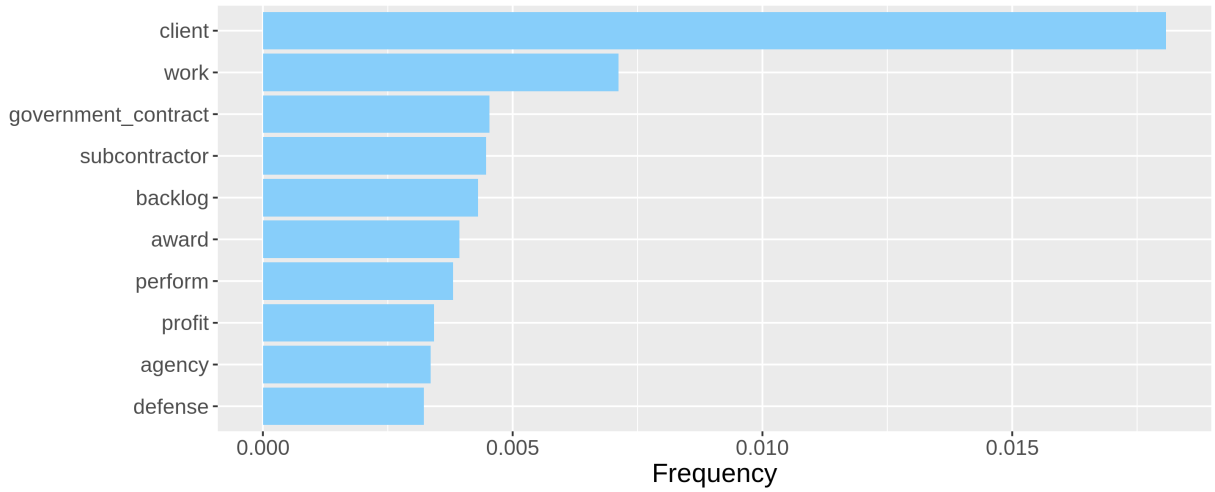
Appendix 1 All Topics

Figure 24: Topic 1



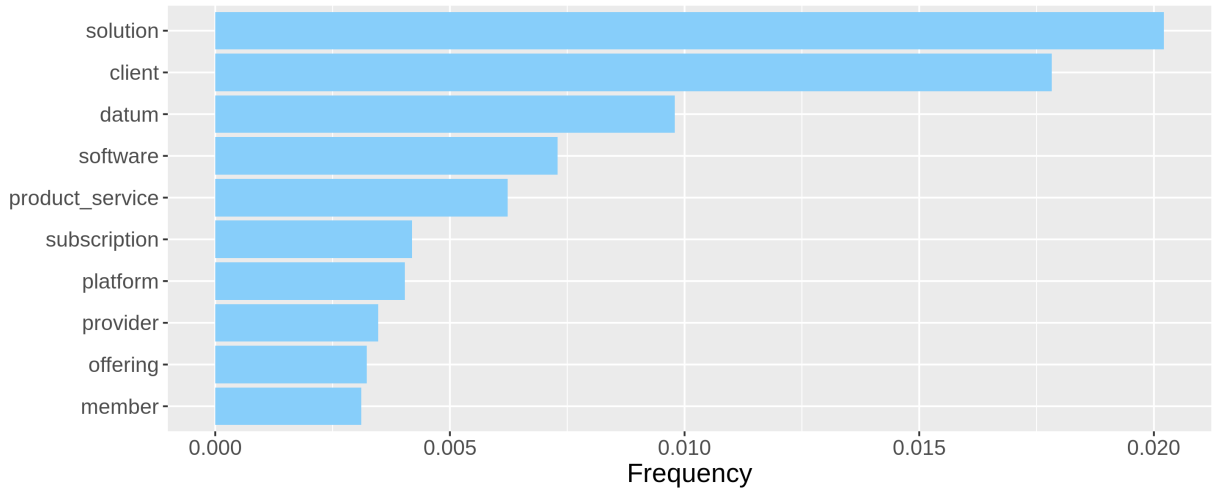
Distribution of the 10 most frequent words for the Topic 1

Figure 25: Topic 2



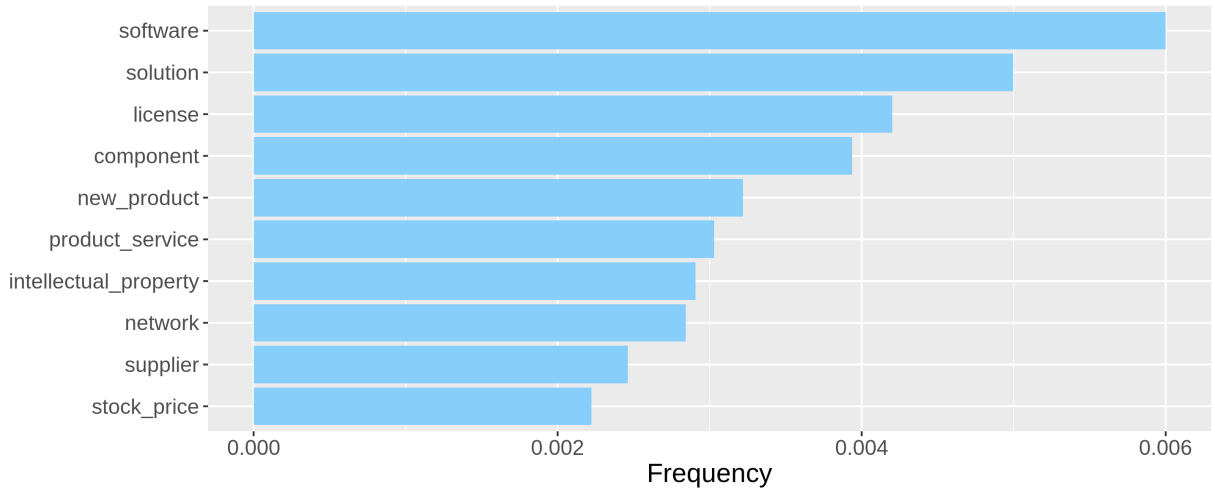
Distribution of the 10 most frequent words for the Topic 2

Figure 26: Topic 3



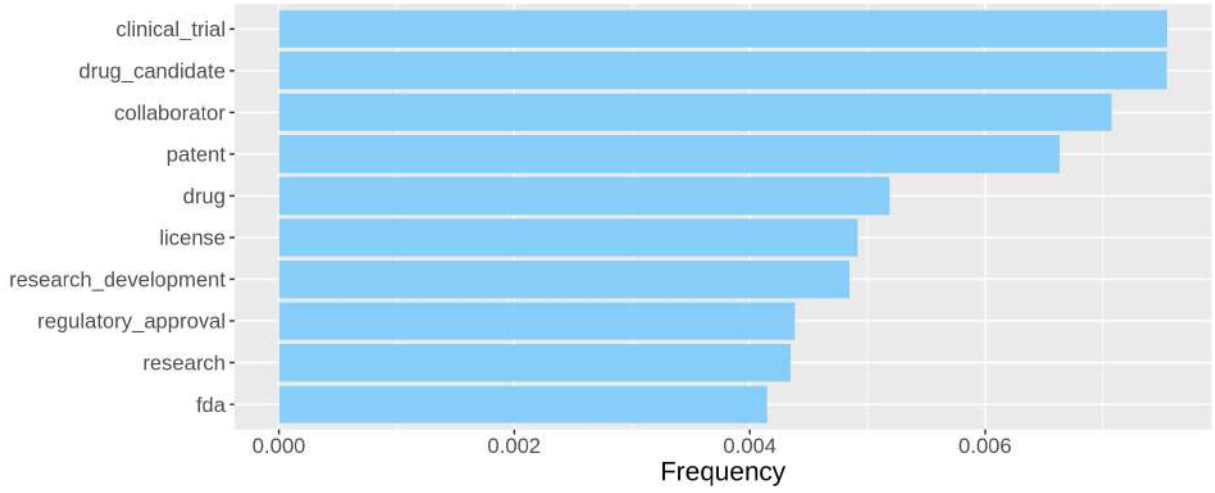
Distribution of the 10 most frequent words for the Topic 3

Figure 27: Topic 4



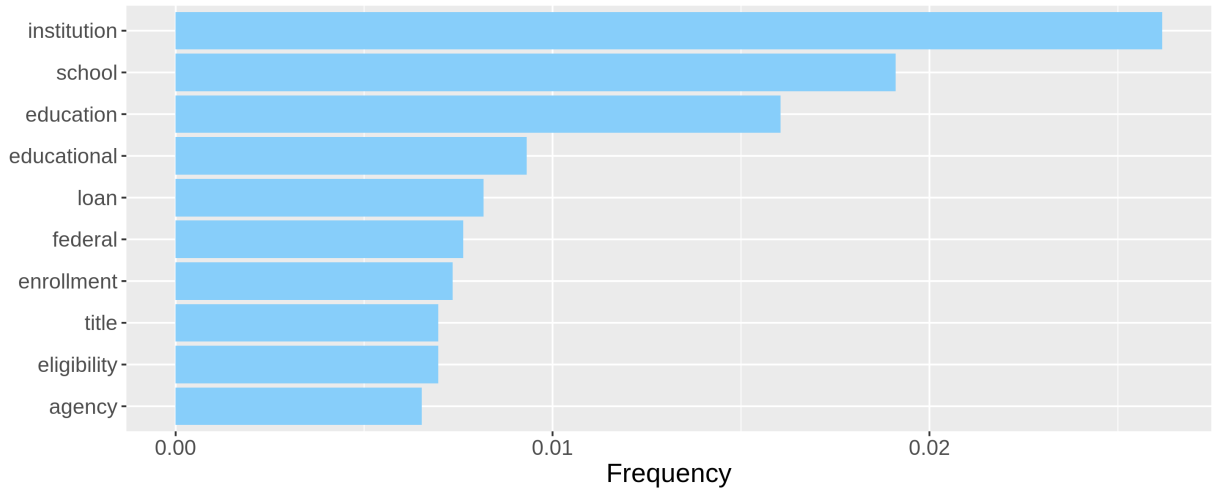
Distribution of the 10 most frequent words for the Topic 4

Figure 28: Topic 5



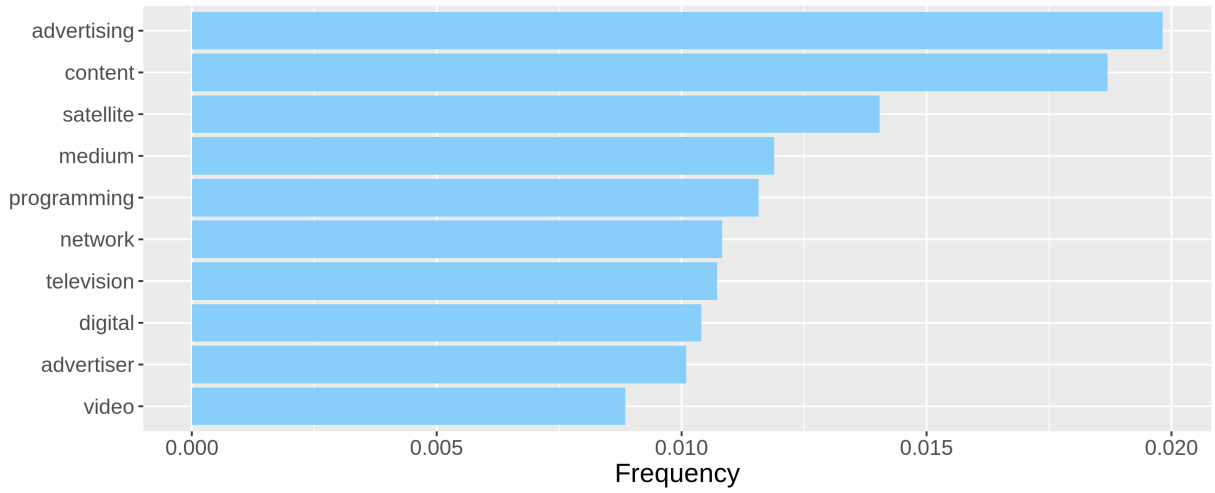
Distribution of the 10 most frequent words for the Topic 5

Figure 29: Topic 6



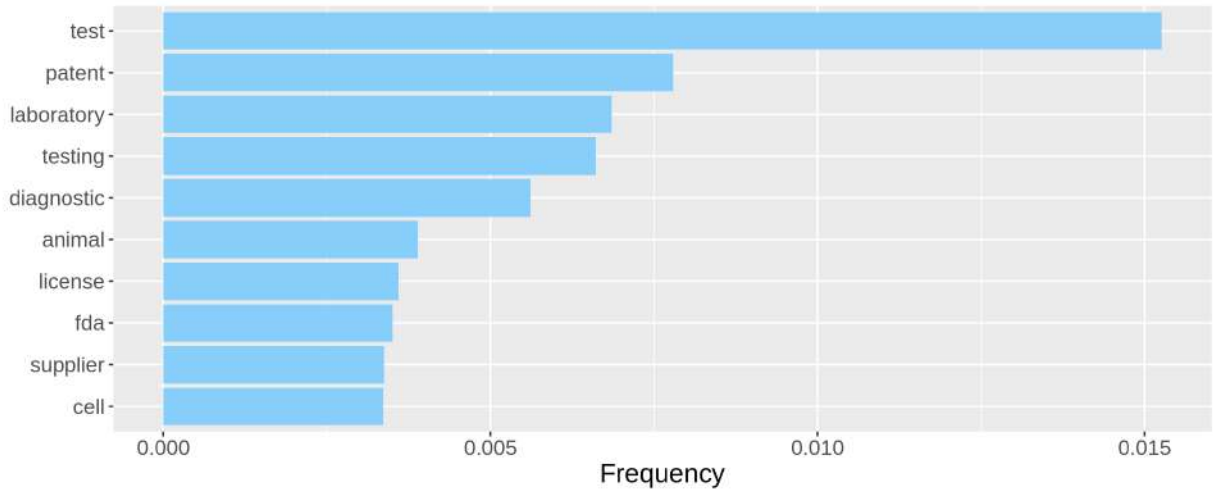
Distribution of the 10 most frequent words for the Topic 6

Figure 30: Topic 7



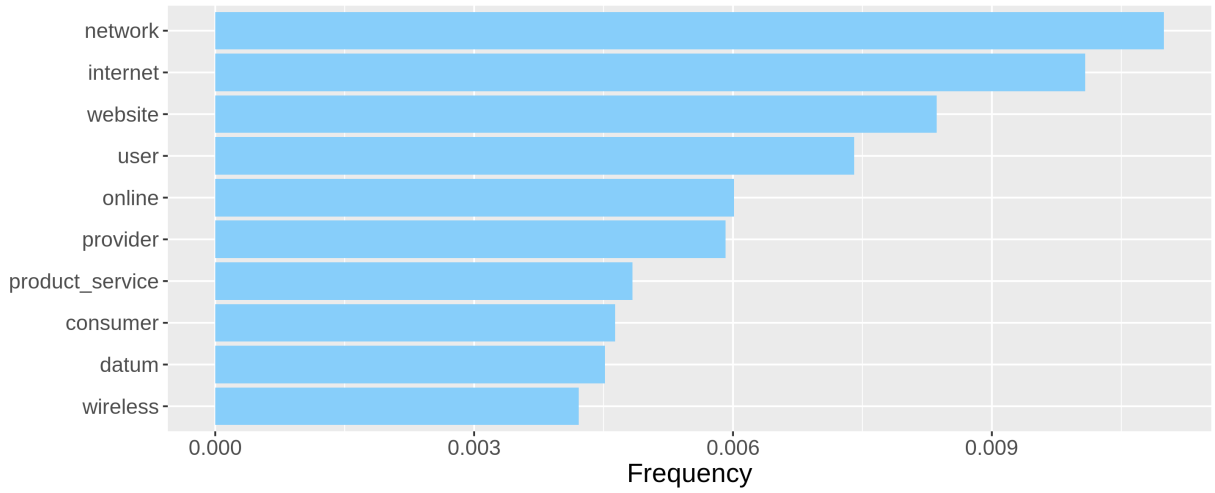
Distribution of the 10 most frequent words for the Topic 7

Figure 31: Topic 8



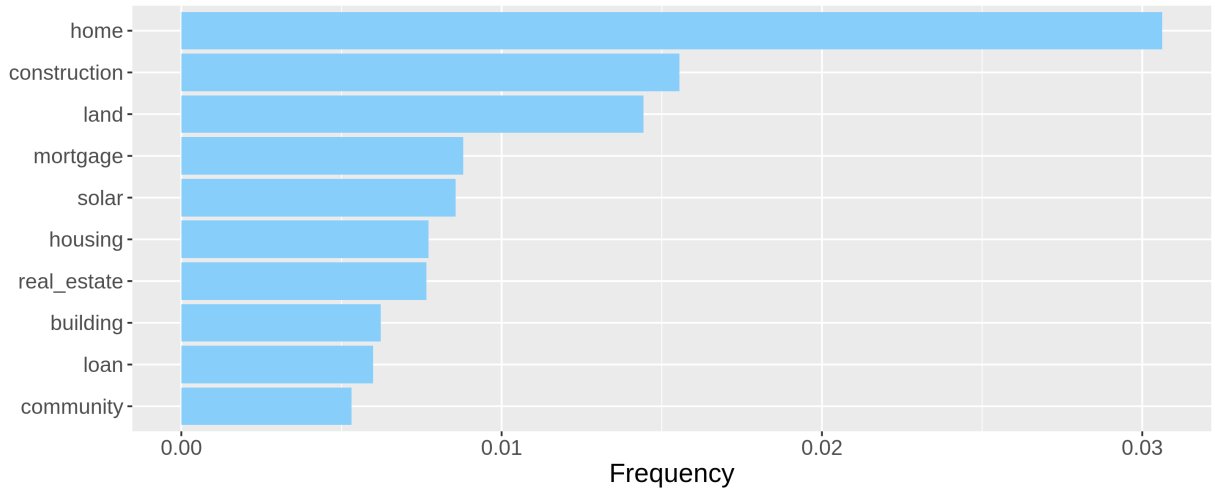
Distribution of the 10 most frequent words for the Topic 8

Figure 32: Topic 9



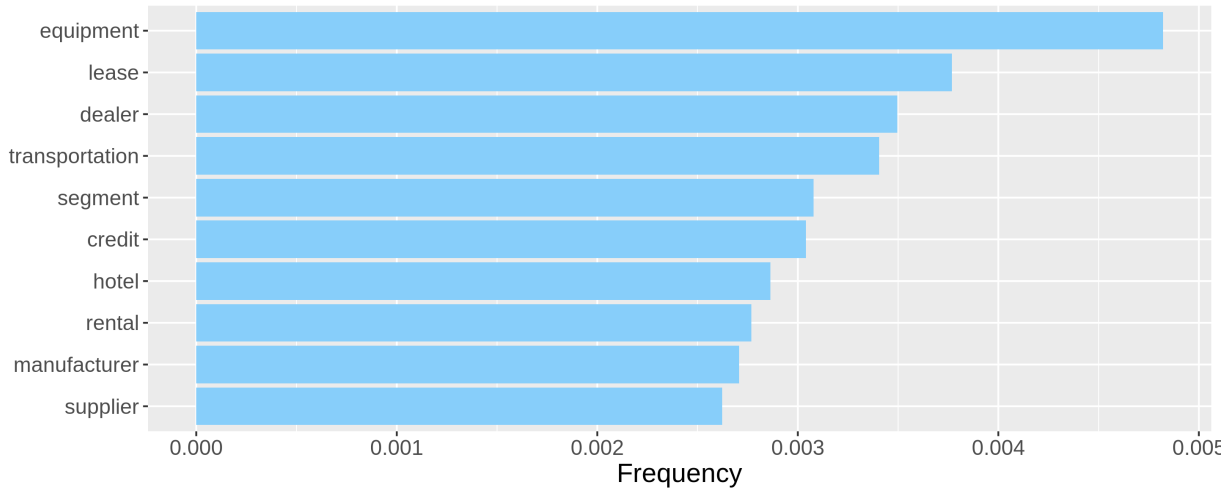
Distribution of the 10 most frequent words for the Topic 9

Figure 33: Topic 10



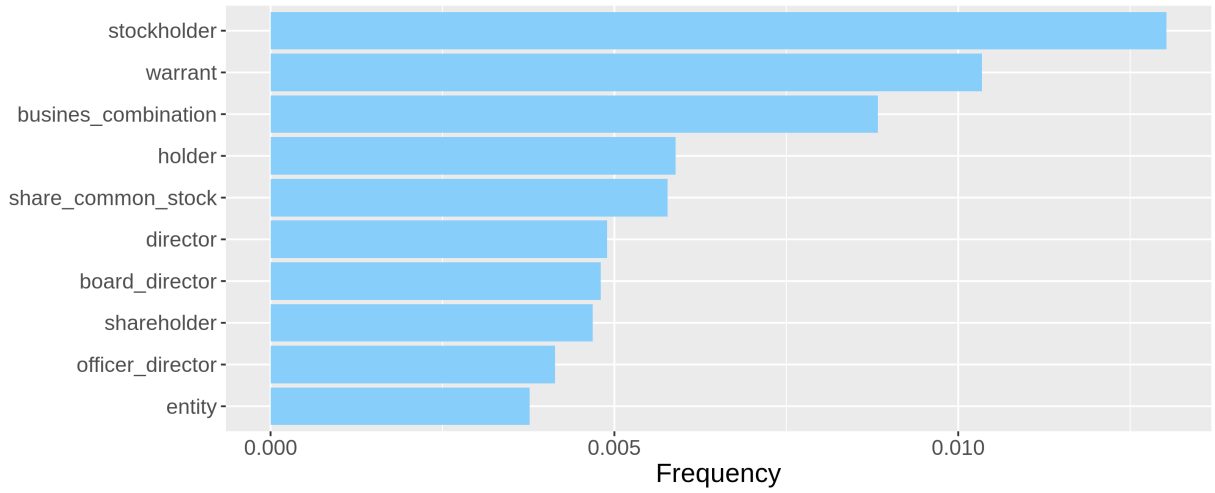
Distribution of the 10 most frequent words for the Topic 10

Figure 34: Topic 11



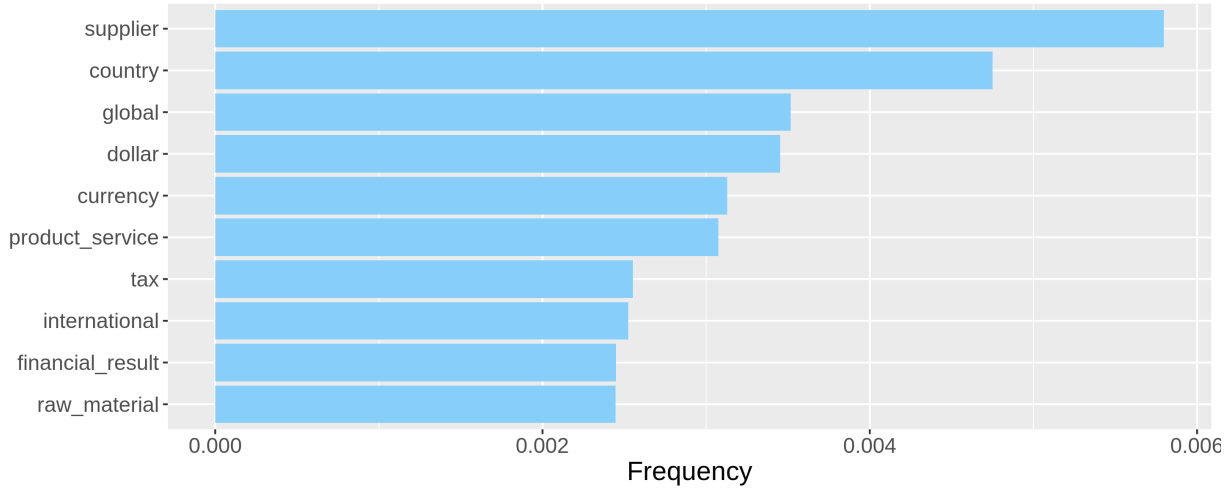
Distribution of the 10 most frequent words for the Topic 11

Figure 35: Topic 12



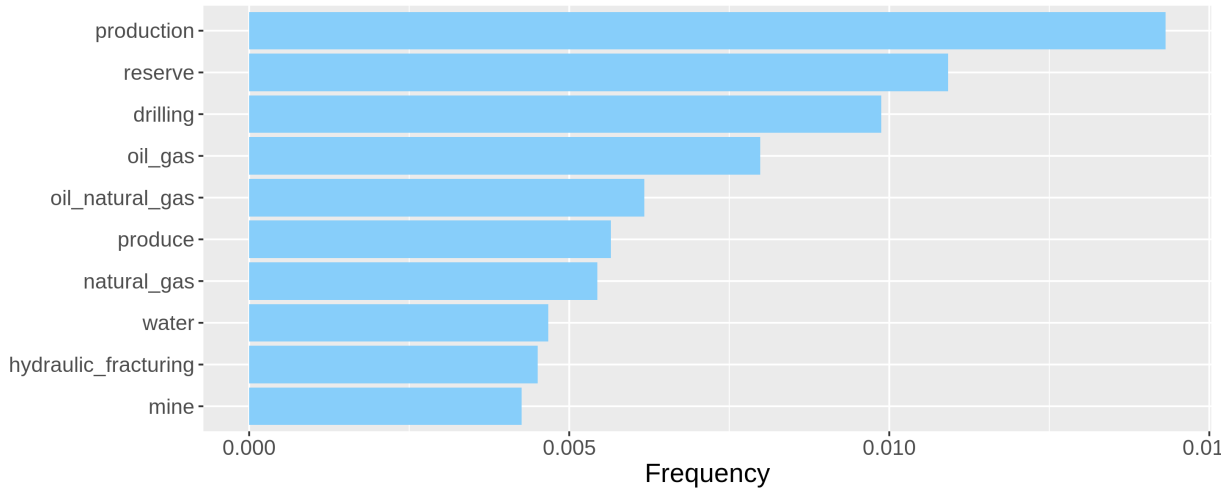
Distribution of the 10 most frequent words for the Topic 12

Figure 36: Topic 13



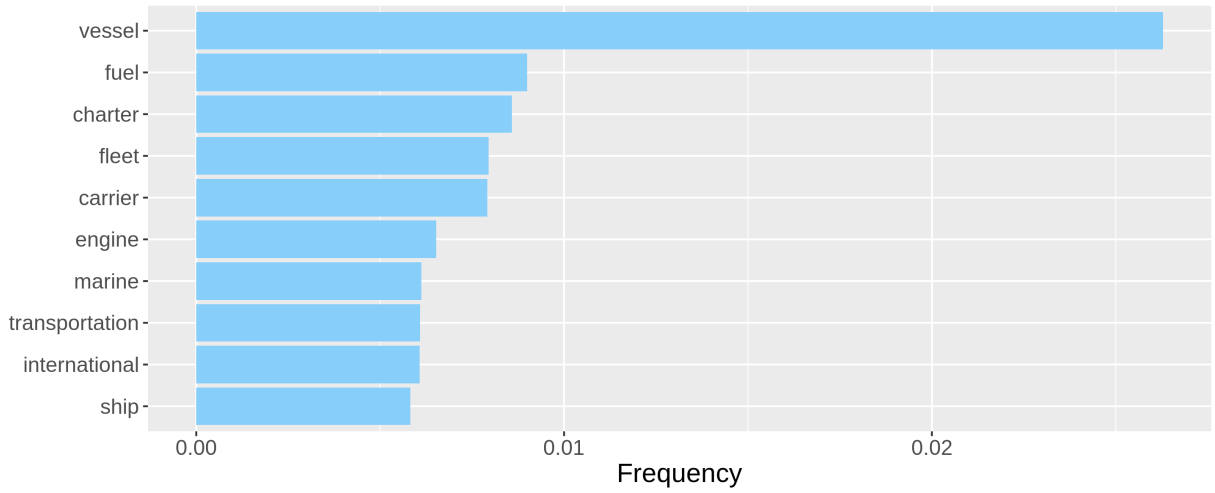
Distribution of the 10 most frequent words for the Topic 13

Figure 37: Topic 14



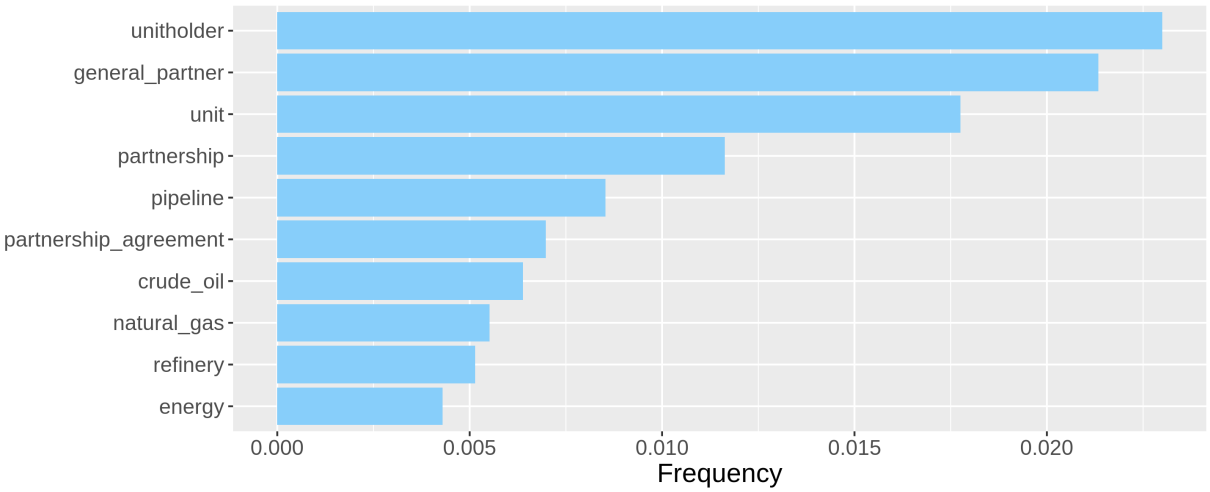
Distribution of the 10 most frequent words for the Topic 14

Figure 38: Topic 15



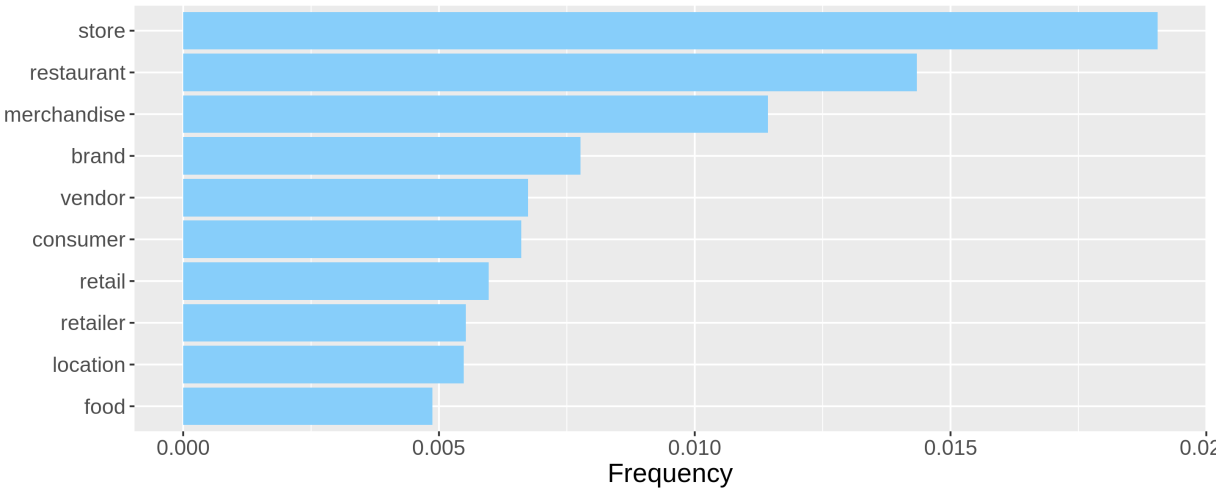
Distribution of the 10 most frequent words for the Topic 15

Figure 39: Topic 16



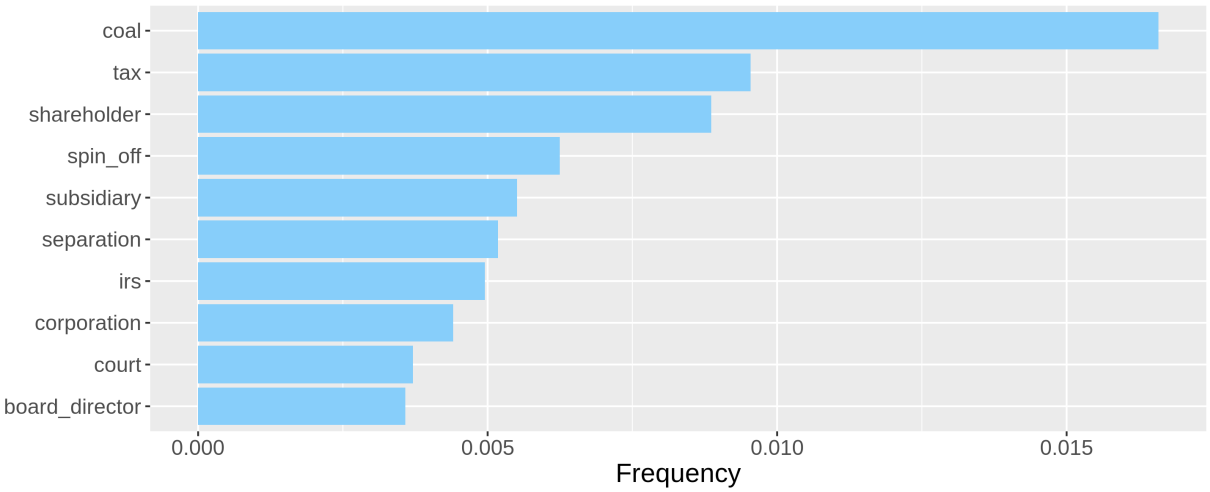
Distribution of the 10 most frequent words for the Topic 16

Figure 40: Topic 17



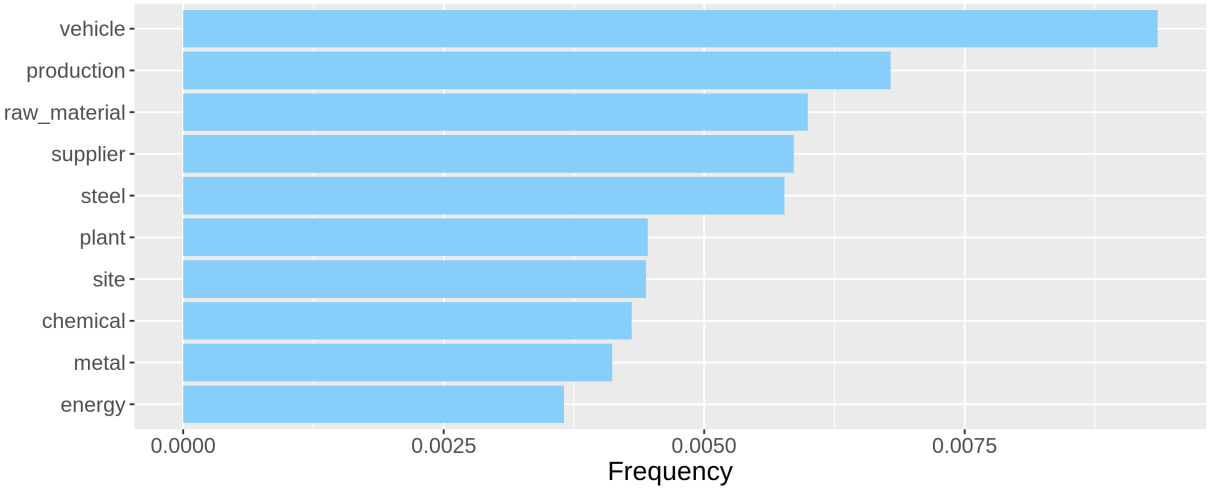
Distribution of the 10 most frequent words for the Topic 17

Figure 41: Topic 18



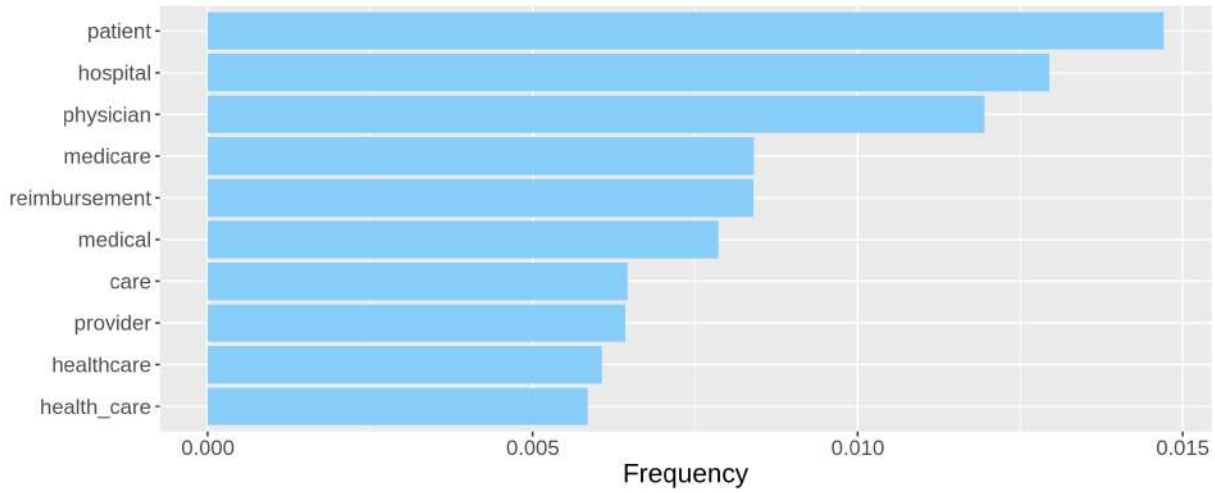
Distribution of the 10 most frequent words for the Topic 18

Figure 42: Topic 19



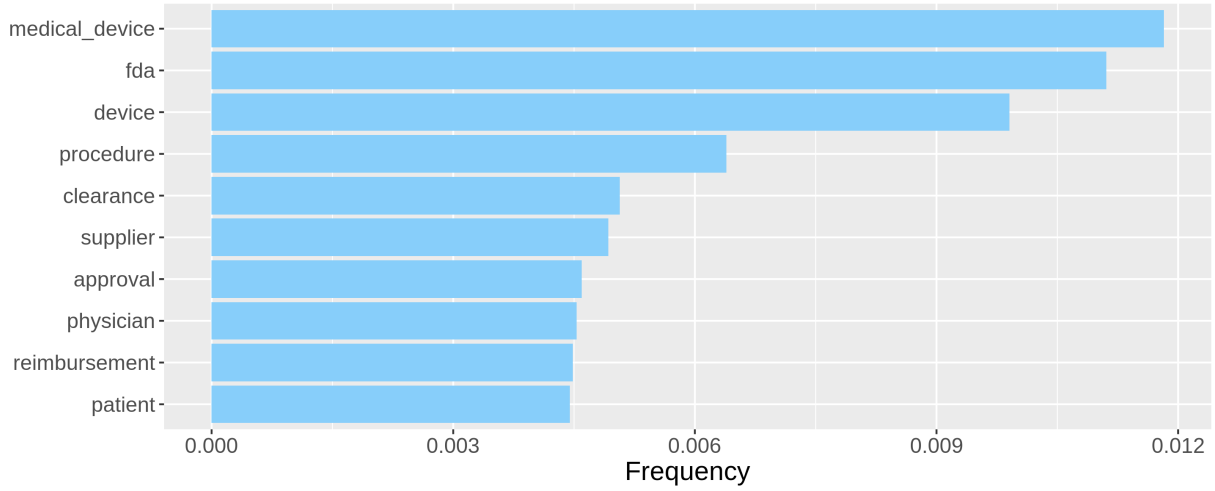
Distribution of the 10 most frequent words for the Topic 19

Figure 43: Topic 20



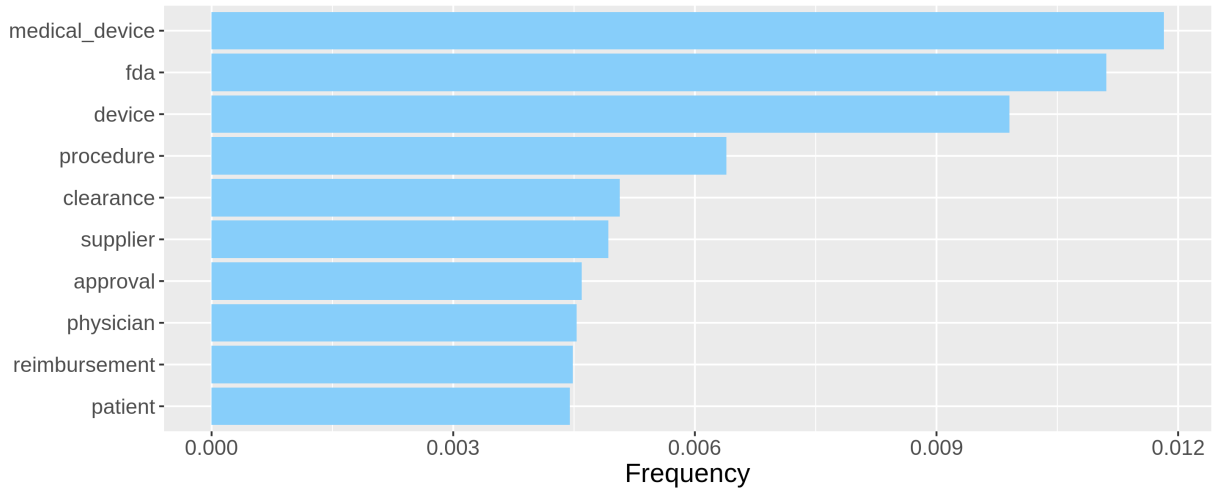
Distribution of the 10 most frequent words for the Topic 20

Figure 44: Topic 21



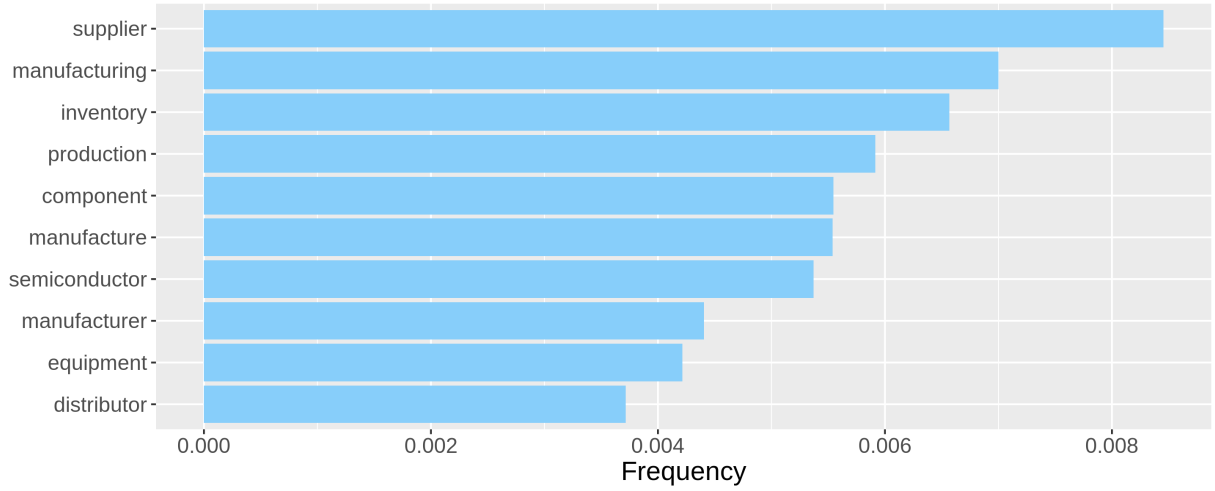
Distribution of the 10 most frequent words for the Topic 21

Figure 45: Topic 21



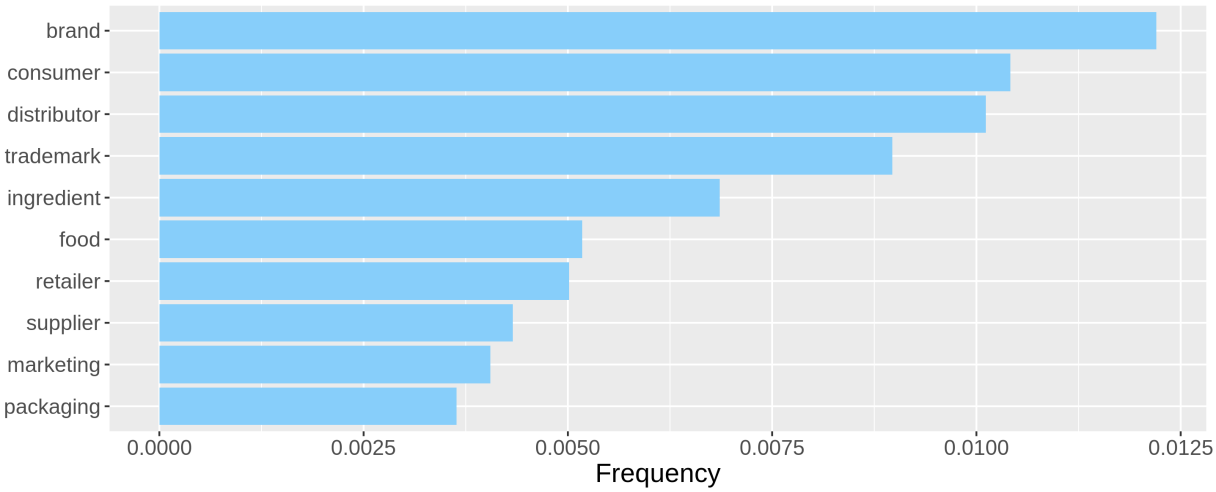
Distribution of the 10 most frequent words for the Topic 21

Figure 46: Topic 22



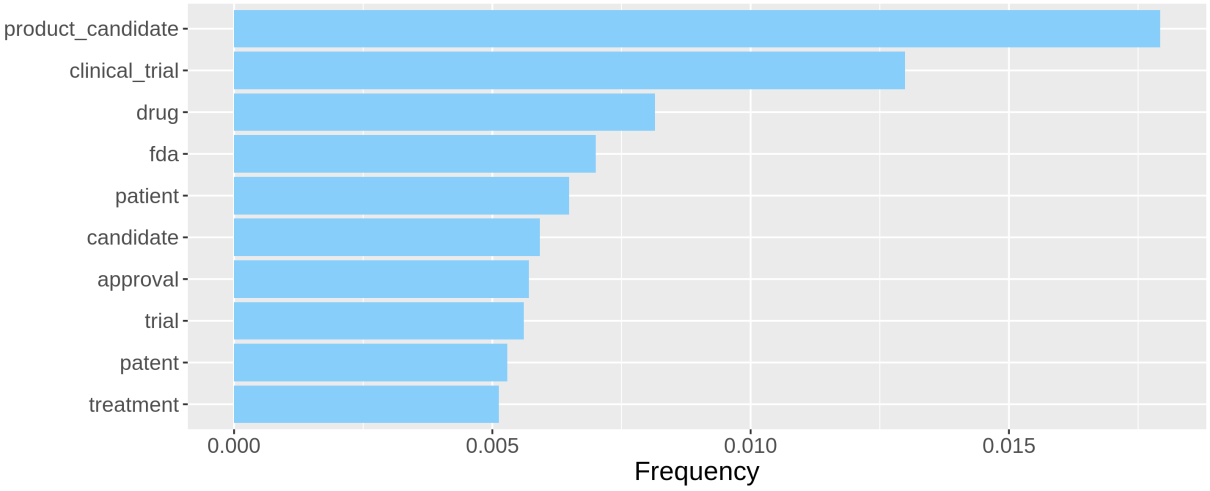
Distribution of the 10 most frequent words for the Topic 22

Figure 47: Topic 23



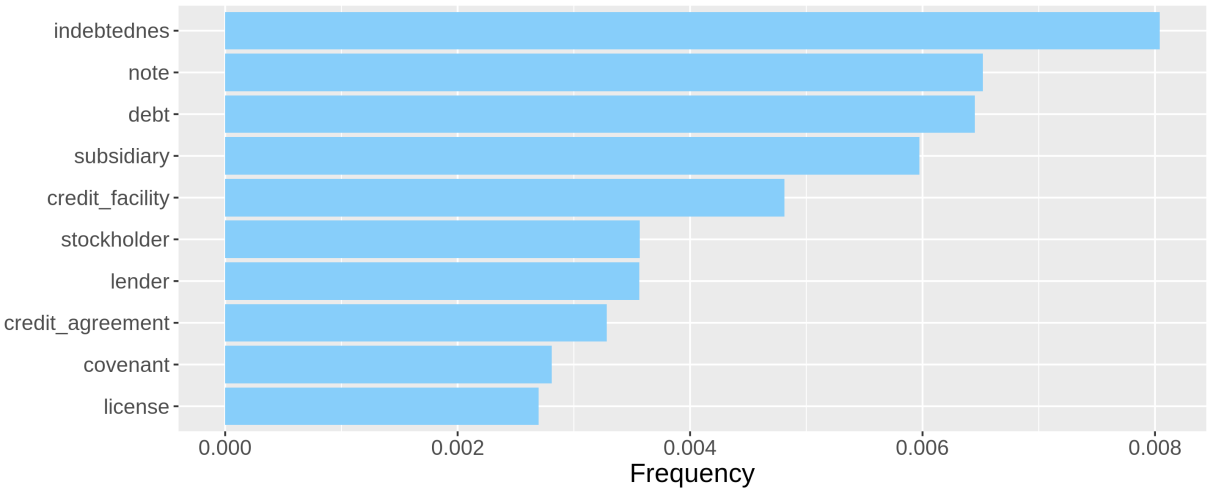
Distribution of the 10 most frequent words for the Topic 23

Figure 48: Topic 24



Distribution of the 10 most frequent words for the Topic 24

Figure 49: Topic 25



Distribution of the 10 most frequent words for the Topic 25

Appendix 2

Figure 50: Correlation of the orthogonal factors with respect to FF5

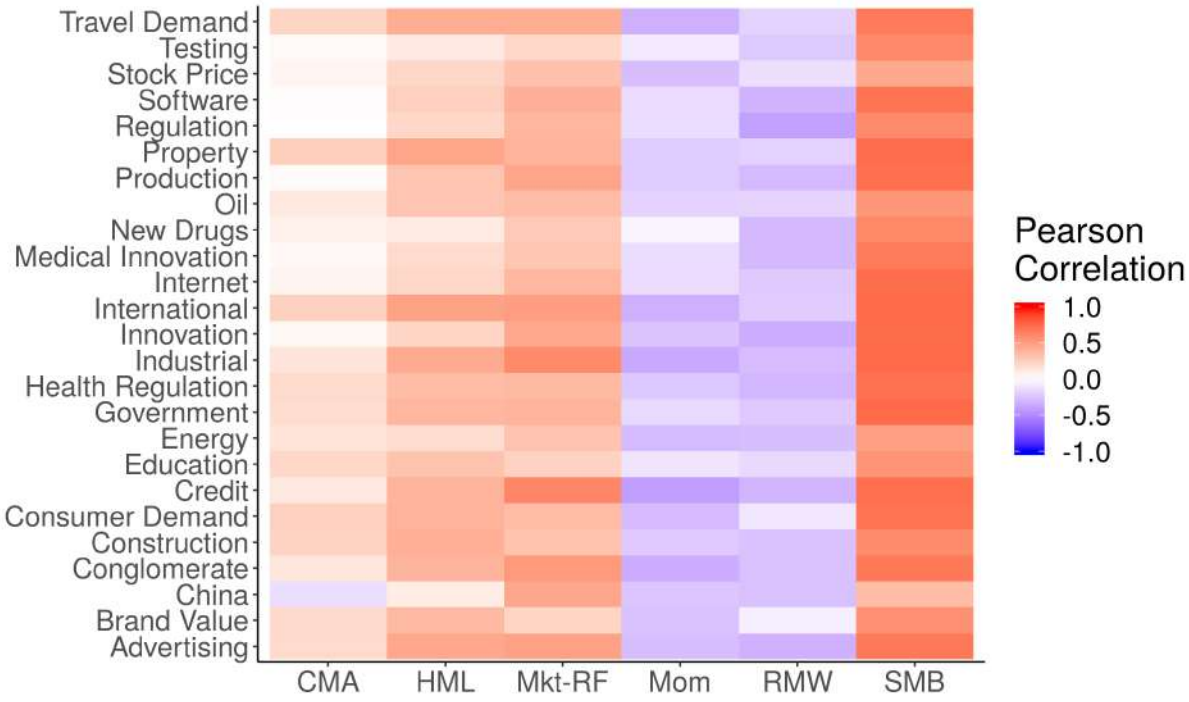


Figure 51: Betas of the orthogonal factors with respect to FF5

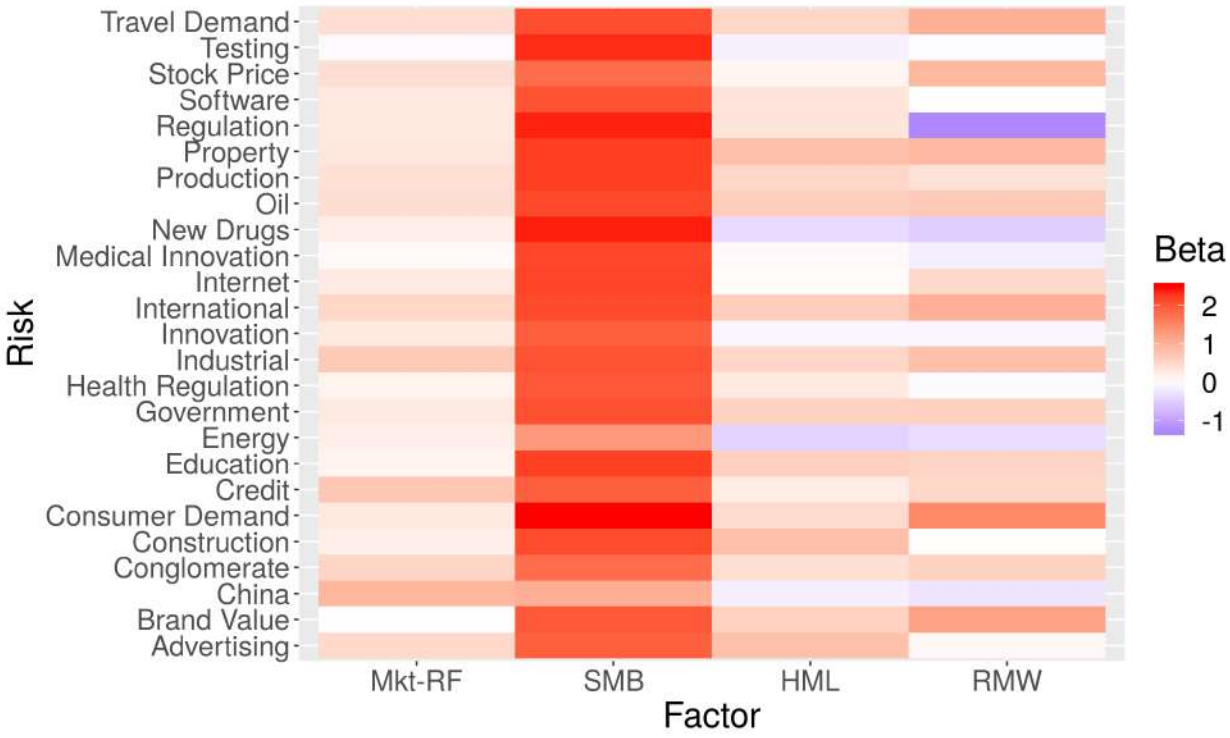


Table 15: Descriptive statistics

Statistic	N	Mean	St. Dev.	Pctl(25)	Pctl(75)
Pairwise Correlation	3,347,132	0.20	0.15	0.10	0.30
Risk Simmilarity	3,347,132	0.14	0.14	0.03	0.20
Beta Exposure	3,347,132	1.25	0.41	0.97	1.50
Book-to-Market Distance	3,347,132	1.05	3.20	0.17	0.92
Size Distance	3,347,132	2.23	1.69	0.89	3.21

Table 16: Correlation Matrix of Distances and Exposures

	Pairwise Correlation	Risk Similarity	Beta Exposure	Book-to-Market Distance	Size Distance
Pairwise Correlation	1	0.19	0.35	0.06	0.12
Risk Similarity	0.19	1	0.03	0.03	0.06
Beta Exposure	0.35	0.03	1	0.06	0.13
Book-to-Market Distance	0.06	0.03	0.06	1	0.16
Size Distance	0.12	0.06	0.13	0.16	1

Table 17: Impact of risk similarity on correlation

	Dependent variable:
	Pairwise Correlation
Simmilarity	0.202^{***}
	(0.001)
	t = 368.520
	p = 0.000
Constant	0.170^{***}
	(0.0001)
	t = 1,554.111
	p = 0.000
Observations	3,619,868
R^2	0.036
Adjusted R^2	0.036
Residual Std. Error	0.150 (df = 3619866)
F Statistic	$135{,}807.300^{***}$ (df = 1; 3619866)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 18: Impact of risk similarity on correlation

	Dependent variable:		
	Pairwise Correlation		
	(1)	(2)	(3)
Risk Simmilarity	0.230^{***}	0.227^{***}	0.225^{***}
	(0.001)	(0.001)	(0.001)
	$t = 382.444$	$t = 387.722$	$t = 384.344$
	$p = 0.000$	$p = 0.000$	$p = 0.000$
Beta Exposure		0.142^{***}	0.141^{***}
		(0.0003)	(0.0003)
		$t = 455.385$	$t = 430.822$
		$p = 0.000$	$p = 0.000$
Book-to-Market Distance			-0.017^{***}
			(0.0003)
			$t = -58.673$
			$p = 0.000$
Size Distance			-0.0003^{***}
			(0.0001)
			$t = -5.244$
			$p = 0.00000$
Constant	0.176^{***}	0.019^{***}	0.027^{***}
	(0.0001)	(0.0004)	(0.0004)
	$t = 1,384.226$	$t = 52.600$	$t = 60.169$
	$p = 0.000$	$p = 0.000$	$p = 0.000$
Observations	3,210,796	3,160,469	3,160,469
R^2	0.044	0.103	0.104
Adjusted R^2	0.044	0.103	0.104
Residual Std. Error	0.158 (df = 3210794)	0.153 (df = 3160466)	0.153 (df = 3160464)
F Statistic	$146{,}263.400^{***}$ (df = 1; 3210794)	$181{,}465.000^{***}$ (df = 2; 3160466)	$91{,}717.650^{***}$ (df = 4; 3160464)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 19: Impact of risk similarity on correlation

	Dependent variable:
	Pairwise Correlation
topic_25	0.081^{***}
	$t = 47.271$
topic_24	0.029^{***}
	$t = 14.539$
topic_23	-0.167^{***}
	$t = -71.966$
topic_22	0.124^{***}
	$t = 109.808$
topic_21	-0.022^{***}
	$t = -8.328$
Observations	3,347,132
R^2	0.114
Adjusted R^2	0.114
Residual Std. Error	0.142 (df = 3347106)
F Statistic	$17{,}236.150^{***}$ (df = 25; 3347106)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 20: Impact of risk similarity on correlation

	Dependent variable:
	Pairwise Correlation
topic_20	0.072^{***}
	$t = 23.205$
topic_19	0.263^{***}
	$t = 198.443$
topic_18	0.219^{***}
	$t = 50.743$
topic_17	0.109^{***}
	$t = 90.658$
topic_16	0.131^{***}
	$t = 30.876$
Observations	3,347,132
R^2	0.114
Adjusted R^2	0.114
Residual Std. Error	0.142 (df = 3347106)
F Statistic	$17{,}236.150^{***}$ (df = 25; 3347106)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 21: Impact of risk similarity on correlation

	Dependent variable:
	Pairwise Correlation
topic_15	0.239^{***}
	$t = 49.889$
topic_14	0.268^{***}
	$t = 148.471$
topic_13	0.366^{***}
	$t = 448.876$
topic_12	-0.673^{***}
	$t = -184.596$
topic_11	0.144^{***}
	$t = 105.497$
Observations	3,347,132
R^2	0.114
Adjusted R^2	0.114
Residual Std. Error	0.142 (df = 3347106)
F Statistic	$17{,}236.150^{***}$ (df = 25; 3347106)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 22: Impact of risk similarity on correlation

	Dependent variable:
	Pairwise Correlation
topic_10	0.464^{***}
	$t = 91.181$
topic_9	-0.062^{***}
	$t = -30.916$
topic_8	-0.181^{***}
	$t = -38.081$
topic_7	0.073^{***}
	$t = 21.495$
topic_6	-0.148^{***}
	$t = -13.518$
Observations	3,347,132
R^2	0.114
Adjusted R^2	0.114
Residual Std. Error	0.142 (df = 3347106)
F Statistic	$17{,}236.150^{***}$ (df = 25; 3347106)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 23: Impact of risk similarity on correlation

	Dependent variable:
	Pairwise Correlation
topic_5	-0.150^{***}
	$t = -64.097$
topic_4	0.030^{***}
	$t = 28.607$
topic_3	0.032^{***}
	$t = 13.251$
topic_2	0.102^{***}
	$t = 68.687$
topic_1	-0.218^{***}
	$t = -44.668$
Constant	0.176^{***}
	$t = 1,451.797$
Observations	3,347,132
R^2	0.114
Adjusted R^2	0.114
Residual Std. Error	0.142 (df = 3347106)
F Statistic	$17,236.150^{***}$ (df = 25; 3347106)
Note:	*p<0.1; **p<0.05; ***p<0.01