Leveraging Information in 10-K Filings to Analyze Risk Trends

By: Michael Chen, Dylan Chou, Spoorthi Jakka, Alden Pritchard

Project Supervisor: Peter Freeman *Project Advisors:* Lars-Alexander Kuehn, Jinghong Liang

Abstract and Introduction

Annual U.S. Securities and Exchange Commission (SEC) reports can come in the form of a 10-K that can contain useful economic indicators in their risk sections: 1A, or "Risk Factors", 7, or "Management's Discussion and Analysis of Financial Condition and Results of Operations", and 7A, or "Quantitative and Qualitative Disclosures About Market Risk". Analysis of these indicators enables industry-specific risk assessments. In our project, we examine 25 different companies, five companies per category over five categories: technology, furniture, food processing, retail and banking. We propose that using the specified risk sections can answer three questions:

(1) Are the topics among 10-K's in the same category similar based on risk discussion?
(2) Will companies with similar risk sections see similar changes in their stock returns?
(3) How does risk in company 10-K reports correlate with stock return risks?

Data Retrieval

We extracted 10-K's in HTML format from the EDGAR database of SEC filings.

Our procedure

(1) Retrieve quarterly tab-separated files from the EDGAR index.
(2) Read in the relevant quarterly 10-K rows per company.
(3) Obtain html files by URL.
(4) Extract sections and store within parquet files.
(5) Identify any sections not picked up by extractor and add those texts to specific rows.
(6) Clean the text via splitting by paragraph, splitting apart multiple different words clumped together after parsing, removing punctuation and extracting only alphabetical characters.

Data Structure

Company	Year	Month	Day	1A	7A	Similar Risk Sections
"Name"	10-K Year	10-K Month	10-K Day	Text if found or NA	Text if found or NA	Text if found or NA

Methodology

We use Python to perform the data retrieval and data cleaning, which carries over to work in Google Colaboratory conducting the following analyses:

(1) Topic modeling using Latent Dirichlet Allocation to analyze contents of specific sections, tracking changes within and between firms over time. LDA achieves this by randomly assigning words to one of a specified number of topics and then counting the frequency of words within each topic.
(2) Creating embeddings of companies' 10-K's using a Doc2Vec model, which transforms similar texts into vectors close together in space, and performing K-means clustering to investigate changes in textual similarity among reports across time.
(3) Combining our text data with monthly stock return data collected from the Center for Research in Security Prices (CRSP) and the annual Compustat-Capital IQ database by the Wharton Research Data Services. We then examine both, sampling data across all years and data for fixed years, for all years. Using this subsetted data, we calculate pairwise cosine similarities and normalize compression distances for our data. For each pair of companies, we also look at the correlation of 12-months of their monthly stock returns and the difference between their annual stock return. We then look for a relationship between our textual similarity metric and our stock returns.
(4) Comparing volatility measures between companies with frequent risk mentions in their 10-K's and those with fewer risk mentions to observe correlations between risk of stock returns and risk mentioned in 10-K reports. Using the previous year's 10-K report risk sections, we rank the 25 firms in order of risk mentions and form high/low risk portfolios among the top five and bottom five companies. For each of these two portfolios, their current year's stock returns volatility is then plotted.

Results

(1) Using LDA for topic modeling finds that risk topics varies widely across companies, though in some years there are topics that are common to most firms. For example, mentions of derivatives are generally low except for a spike in 2009, directly after the financial crisis, as seen in the ensuing topic graph. **Topics for individual firms generally are similar in consecutive years and became more dissimilar as the time between years increases.**

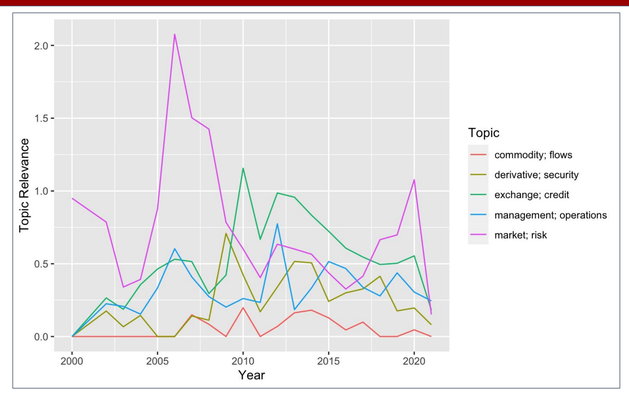

Figure 1. Topic Trends across Firms

Results (cont.)

(2) Calculating the cosine distance between embeddings for each report reveals that all reports within the same year tended to be similar, and it appears that **firms from the same industry are not assigned to the same cluster any more often than firms from different industries**. Below we plot embeddings of text from each firm's 10-K across all years colored by K-means cluster(left) and the firm's industry(right).

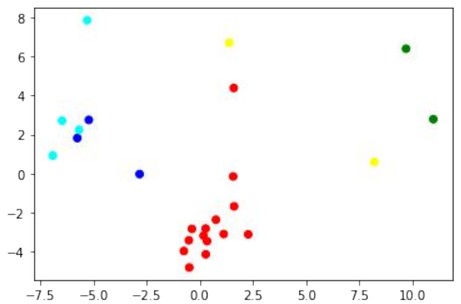

Figure 2. Company 10-K's colored by K-means

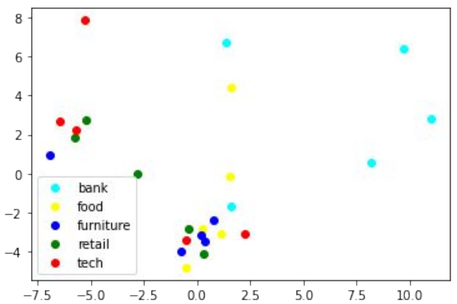

Figure 3. Company 10-K's colored by Industry

(3) We look at the correlation between annual stock returns and the difference between annual stock returns for companies both across years and for fixed years as a measure of similarity in the changes of their stock returns. **We do not detect a significant relationship between the similarity of two companies' risk sections, for sections 7A, 1A, and combined, and their stock returns.**

(4) Prior to the Stock Market Crash in 2008, **the standard deviation of the stock return % changes, or stock return % volatility, for firms with high risk mentions in their 10-K reports is greater than that of low risk firms, but after 2008, low risk firms have greater volatility than that of high risk firms**, as shown in Figure 4. This corresponds to more high risk companies after 2008 being in the food industry, which may have lower stock volatility shown in Figure 5. More low risk companies after 2008 are in furniture shown in Figure 6, which could be more volatile due to changing prices in raw materials or a reduced workforce due to COVID-19.

Figure 4. Volatility of High/Low Risk Companies

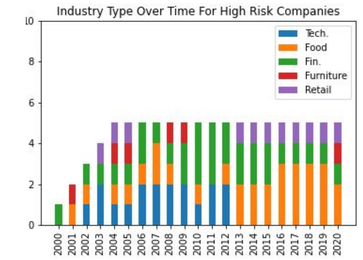

Figure 5. High Risk Industries

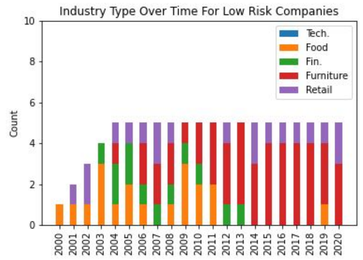

Figure 6. High Risk Industries

Conclusions and Discussion

Our analysis finds that most 10-K documents use similar language. This enabled us to track trends over time and see when macro events causes firms to focus on specific business areas

Groupings of portfolios by risk mentions in 10-Ks reveals that a rise in food companies with high risk 10-Ks and more furniture companies with low-risk 10-Ks leads to a reverse in volatility of high and low risk portfolios post-2008. This swap could be due to a greater concern of food contamination that would make food companies with low volatile stocks have high risk 10-Ks. Furniture companies could have lower risk 10-Ks, but highly volatile stock prices due to changing raw material prices or COVID-19 warranted workforce fluctuations. To properly identify valid factors driving the change in 10-K risk or volatility, we would need to conduct further research based on economic events.

Our data engineering process was hampered by the different HTML formatting of 10-K documents, so future work can be done in handling different company 10-K and automating collection to analyze more firms. We also do not detect any significant relationship between our similarity metrics and our stock returns. This could be due to noise in the data and in the future we could try to group companies together into portfolios, possibly using the K-means groups we found, and run a portfolio-level analysis.

References

Kang, Taeyoung, Do-Hyung Park, and Ingoo Han. "Beyond the numbers: The effect of 10-K tone on firms' performance predictions using text analytics." Telematics and Informatics 35.2 (2018): 370-381.

Stock return data from: https://wrds-www.wharton.upenn.edu/pages/about/data-vendors/center-for-research-in-security-prices-crsp/