Can Machines Learn Finance?
–Textual Analysis of 10-K MD&A

Zhou Xing* Chia-Yun Chang † Songrun He ‡

March 19, 2021

Abstract

We conducted comprehensive content analysis on the Management Discussion and Analysis section of companies' annual report. We applied word count, topic modeling and word embedding methods to explore the industrial organization patterns, informational content related to stock returns and managerial visions and values. From word count, we find clear industrial organizational patterns and highly meaningful words associated with negative stock returns. We quantify the aggregate stock market sentiment using LM dictionary and find that the year 2020 has a significant impact on the real economy. From topic modeling, we were able to find latent topics both human-decipherable and undecipherable. The topic attention changes are used for stock return prediction. We observe the potential of the semantic features for stock return prediction. For the word embedding part, though the word2vec model did not provide significant structure of the embedding vector space, it gave intuitive results in mapping key words to the risk-uncertainty and profit-social responsibility dimensions.

Keywords: Content Analysis, Management Discussion & Analysis

*The University of Chicago, Division of Social Science. zhouxing@uchicago.edu
†The University of Chicago, Division of Social Science. cchiayun@uchicago.edu
‡The University of Chicago, Division of Social Science. hesongrun@uchicago.edu

1 Introduction

> *"... One of the most promising use of relatively new AI techniques may be processing unstructured natural language data in the form of news articles, company reports and social media posts, in an effort to gain insights into the future performance of companies, currencies, commodities, or financial instruments. ..."*
>
> – MIT Technology Review

In this study, we carried out a comprehensive content analysis on the Management Discussion and Analysis section (MD&A) of companies' annual report. The content of MD&A ranges from business status, SWOT analysis, financial situations, performance analysis, threats and opportunities. It is one salient element that provides insightful information on the performance of the company in view of various macro-economic barriers under which it operates.

We hope to harness the great power of content analysis address many important issues in economics and finance: how do stock market sentiment change over time, what are the industrial organizational patterns of companies, how do managers assess risks versus uncertainty(the famous dichotomy proposed by [5]) and how does the rise of ESG investing shift the focus of corporations between profits and social responsibilities. We also trained word2vec models over corpus for industries, and investigate the heterogeneity across the four sectors. Many of our results are far from conclusive but we believe content analysis provides an unique perspective that are highly intuitive and interpretable.

There have been many applications applications of textual analysis in economics and finance. [7] pioneered the research in quantifying the market sentiment using wall street journals. [6] produces an influential financial dictionary that captures the unique context of financial settings. [2] quantify the overall manager sentiment with earnings call transcript and financial reports. They found significant predictive power of their manager sentiment index to the aggregate stock market.

However, most of these analysis are based on word counts. This tradeoff is shown in figure 1. The counting methods have great economic interpretability and computational scalability. However, it ignores the linguistic complexity in the

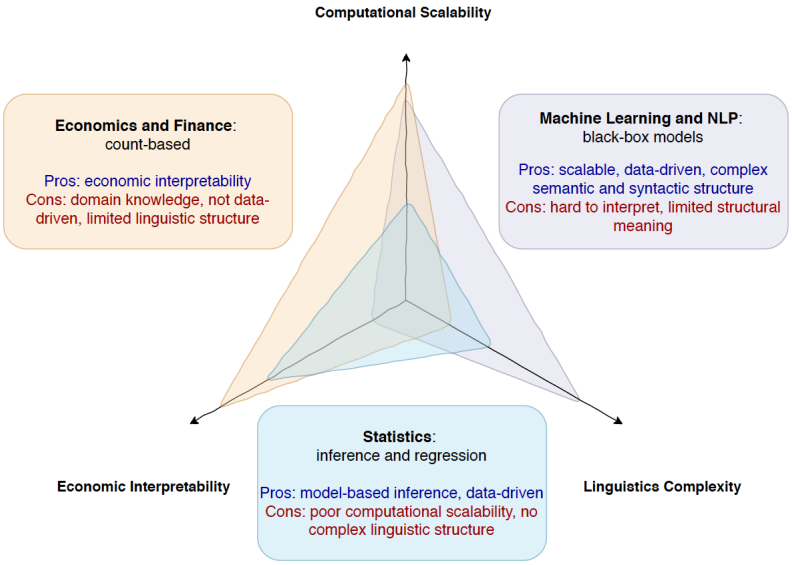

Figure 1: Different Approaches, this figure comes from [1]

underlying language.

In this study, we not only analyzed the corpus with word counts, we also construct topic models and explored the semantic space with word embeddings. These methodology provides much richer insight into our corpus.

Firstly, from word counting, we find significant industrial clustering phenomenon (considering tf-idf weighting). There are distinctive word distribution patterns across different industries. We also link the documents with stock returns and rank the word by their co-occurrence with positive words. We find that the negative word list are highly meaningful. Furthermore, we quantify the overall market sentiment using Loughran and McDonald dictionary and find that in the year 2020, there is a surge in negative words, uncertainty words, weak modal words and a decline in positive words. This suggests covid-19 has a significant shock to the real economy.

For the topic modeling section, we first visualized the term frequency vectors of the annual reports using PCA and found clear clustering. This prompt us to use K-means clustering on the data to find sector related topic words. Secondly, we used LDA topic modeling on the corpus. We conducted 4 experiments using LDA, each with different data-prepossessing criterion and forms of input. In this section, we are able to capture some human-decipherable topics that align with

the different sectors, and some undecipherable topics that shows latent topic that aligns with the functionality of company annual reports. We then used topic attention change to predict stock return.

The rest of the paper is organized as follows. Section 2 introduces our data. Section 3 lays out our analaysis based on word counts. Section 4 is focused on topic modeling and clustering. Section 5 presents our exploration of the semantic space with word embeddings. Section 6 concludes.

2 Data

2.1 The 10-K Sample

We download all 10-Ks from the EDGAR website (www.sec.gov) over the year 2010 to 2020 with the python API sec-edgar. There are over 80,000 documents in our sample period. We then merge the 10-K sample with CRSP database requiring that the company is traded publicly in the stock market and reported on CRSP as an ordinary common equity firm. After the merge, we have 37,720 documents in total. The drop in number is not surprising as there are many real estate, nonoperating, or asset-backed partnerships/trusts that are not publicly traded but are required to file with SEC. Our final corpus has 5405 companies with 242,263,144 words in total. Figure 2 shows the number of articles over time. We can see that our sample is very balanced across time.

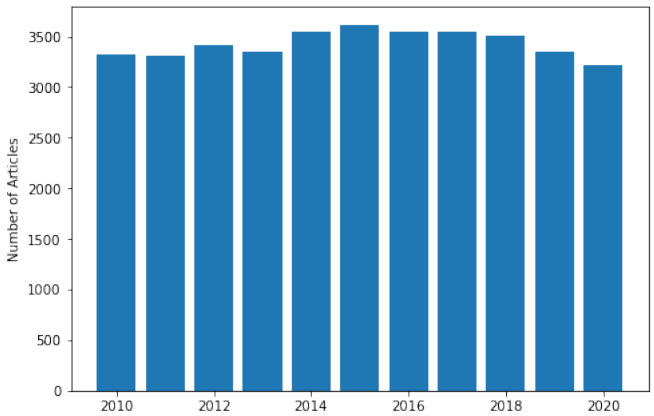

Figure 2: Number of articles over time

Our analysis focus on the Management Discussion and Analysis subsection (item 7) of the 10-K. This is a separate section in the company annual report that provides insightful information on the business status and financial situations of the company in view of the various macro-economic barriers under which it operates. This section contains the most important textual information in 10-K reports.

2.2 Parsing the 10-K MD&A

To parse the 10-K, we first apply beautiful soup to the html files downloaded. After that, we look for item 7 in the html. We start read the management discussion and analysis from item 7 until we see item 8. We exclude extracted text with less than 1000 characters because there are some cases the MD&A section information is 'incorporated by reference'(typically deferring to the shareholders annual report). At text level, we also remove special characters in the text.

After these procedures, we performed data cleanning at the word level. We harness the great power of lucem_illud and performed tokenization, removal of stopwords, and lemmatization to our corpus.

2.3 Stock Returns and Industry Classifications

We also link our corpus with stock returns and company industry classifications. The stock return data are from the CRSP database. We calculated the three-day time window of cumulative return of the company relative to the market benchmark around the announcement date. It can be formalized as follows:

$$(1 + R_{i,t-1})(1 + R_{it})(1 + R_{i,t+1}) - (1 + R_{m,t-1})(1 + R_{mt})(1 + R_{i,t+1})$$

where i denotes stock and m denotes the market. We exclude market return to truly pick up the unique idiosyncratic signal for the company. We believe this return measure can fully reflect price responses to the company annual announcements.

We will adopt the GICS industry classification measure. GICS is a three-layer

industry classification system. In our analysis, to have an overview, we will mainly focus on the first-level classification. We obtain the industry classification from the S&P Compustat database.

3 Word Counts

In this part, we mainly report our analysis based on word-count. This simple technique offers much insight and intuition into our large-scale corpus, which lay a solid foundation for our application of more advanced methods. We first present the wordcloud made from all documents and some specific industries. Next, we link the word with stock returns and screen for word that often co-occurs with positive/negative stock returns. Finally, we quantify the overall stock market sentiment using the Loughran-McDonald financial dictionary proposed by [6].

3.1 Word Clouds

Figure 3 presents the wordcloud made from all documents. We can see there are frequent occurrence of many words in the financial context. This is reasonable as financial aspect will be an important focus for the MD&A section.

Figure 3: Wordcloud from All Documents

Furthermore, using GICS, we filter out four industries: finance, IT, energy and material. We plot the word cloud for each of them separately. During our initial experiments, we used the raw wordcount to produce the wordcloud. However, the results do not look very intuitive with distinctive industry features. There are many financial-related words in all these four industries. To increase the signal, we used the aggregated tf-idf for all the documents of a certain industry.

Figure 4: Word Cloud for Four Industries: Finance(Upper Left), IT(Upper Right), Energy(Lower Left), and Materials(Lower Right)

The resulting wordcloud is shown in figure 4. From this figure, we can see clearly there are distinctive patterns in word distribution across different industries. These words are highly intuitive and we can see immediately the emphasis of business operations behind different industries.

3.2 Sentiment Charged Words

Next, we consider the documents and stock returns jointly. We rank the words by their co-occurrence with positive stock returns. To be specific, we follow the marginal screening procedure proposed in [3]. We calculated the frequency of word j co-occurring with positive stock returns:

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

We then exclude the word occurring less than $2,000$ times in all articles to reduce the noise. After that, we rank all the word by their corresponding f values.

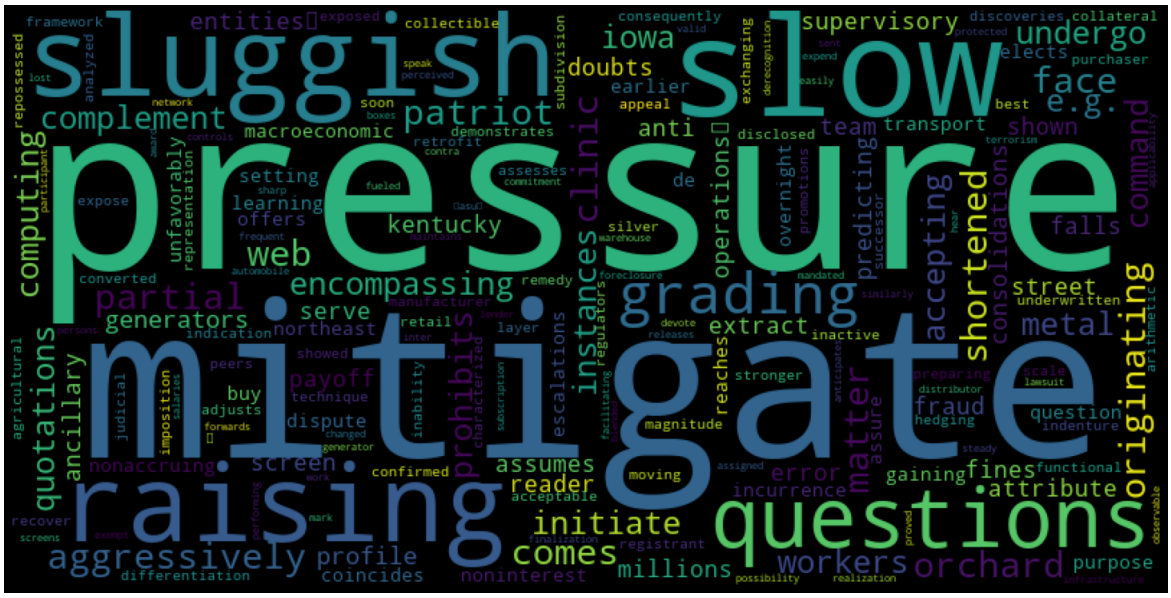

Figure 5: Negative Sentiment Charged Words

Figure 5 presents the top 200 negative words with lowest f-values and figure 6 shows the top 200 positive words with highest f-values. We examined the word list and find that the negative word list is highly intuitive while the positive ones contain much noise. This can also be visualized in figure 5 and figure 6.

After examining a sample of our corpus, we find that the reason behind the difference lies in the 'negations' used by managers. This has also been pointed out in [6]. It is more common for managers to frame negative events using negation plus positive words. It is very rare to see them conveying positive news using negation plus negative words. This asymmetry effect means that negative words are much more informative compared to positive ones.

Figure 6: Positive Sentiment Charged Words

3.3 Overall Stock Market Sentiment through Word Counting

Last but not least, we quantify the overall stock market sentiment using Loughran-McDonald financial dictionary proposed in [6]. This is a dictionary designed to fit the financial context. The authors provide 7 word list: negative, positive, uncertainty, weak modal, strong modal, litigious, and constraining words. In our exercise, we will mainly focus on the first 5 word lists.

There is significant seasonality pattern in number of 10-K documents produced. Most companies tend to have a fiscal year end on 12.31. Therefore, we observe huge spikes in March or April for 10-K reports. To mitigate this, we calculate the 12-month moving average of average term frequencies for all the word lists.

Figure 7 shows the average word count for different word lists over time. We also provide some example words above each figure. From the figure, we can see there is a significant rise in negative words, uncertainty words, weak modal words and a decline in positive words in the year 2020. From this analysis, the managers become very negative about the prospect of the companies at the onset of the covid crisis. This is reflected in the increase in their use of negative words and decrease in their use of positive words. From the use of weak modal words and

uncertainty words, we can see managers become highly uncertain about the future during this time of heightened volatility. The year 2020 seems to be significant shock to the real economy.

To extend on this, we may also try weighting each document by the corresponding market value of the company because the sentiment from larger companies will have greater impact on the overall sentiment of the market.

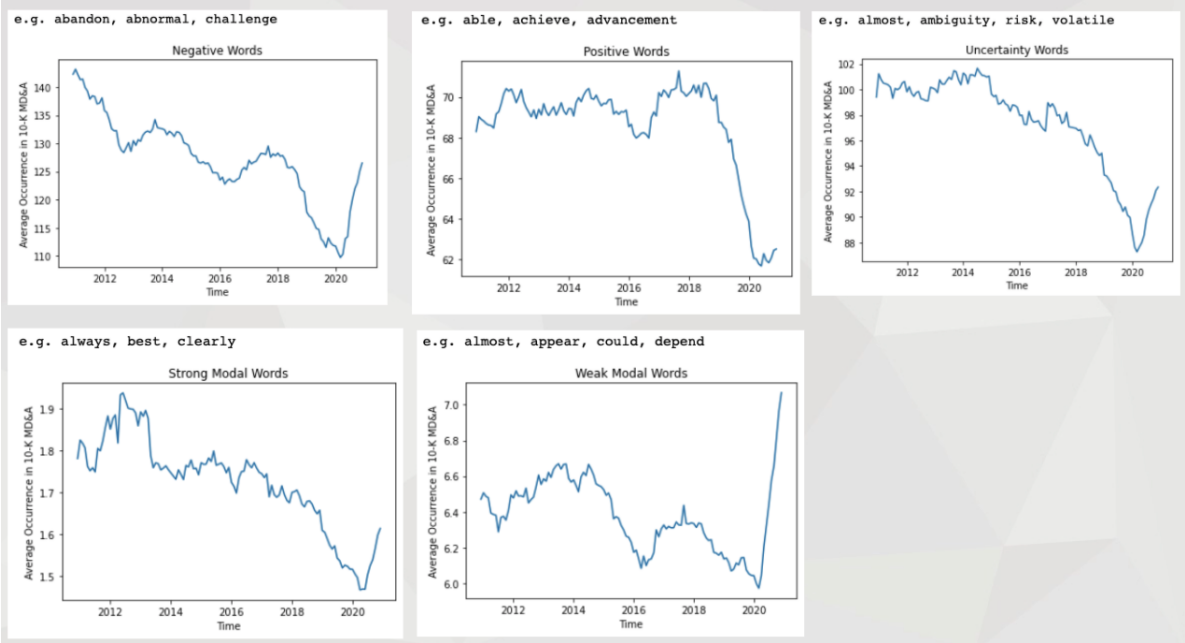

Figure 7: Aggregate Stock Market Sentiment over Time

3.4 Predicting Stock Returns using Word Count

Inspired by the marginal screening results in 3.2, in this part, we hope to investigate whether we can predict stock returns based on the word frequency distribution of the 10-K MD&A section. We consider this as a simple classification problem where there are two categories: positive stock returns and negative stock returns. Different from section 3.2, we are going to consider return at day $t+1$ instead of the three day cumulative returns since we are interested in return prediction.

We consider five setups here:

- use raw word count of all words
- use raw word count based on LM dictionary

- use tf-idf vector of all words
- use tf-df vector of words in LM dictionary
- use the aggregate word counts from 7 of the LM dictionary word lists

For each setup, we consider three prediction algorithms: (1) logistic regression; (2) random forest; (3) multi-layer perceptron. We split the dataset into training set and test set. The training set consists of the documents from 2010-01-01 to 2018-12-31 and the test set consists of the documents from 2019-01-01 to 2020-12-31.

Setup	Logit	RF	MLP
All-count	0.512	0.507	0.496
LM-count	0.516	0.511	0.491
All-tfidf	0.508	0.503	0.494
LM-tfidf	0.526	0.507	0.503
LM-agg	0.547	0.506	0.539

Table 1: Stock Return Prediction: Score in the Test Set

Table 1 shows the prediction score in the test set. From the analysis we can see that given the low signal-to-noise ratio, it is very difficult to predict stock market returns based on the word frequency count of the MD&A section of company annual report. Even the best model using aggregate wordlist from LM dictionary can only achieve a predict score of 0.547 which is only slightly better than guessing at random. This again highlights the micro-efficiency feature of the financial market where the news are reflected in the stock prices immediately after its release. There seems to be no lag in market response for traders to exploit.

4 Topic Modeling Using K-means and LDA

Through word count methods, we have found that annual reports of different sectors do demonstrate vast distinction between their use of words. We hope to harness the distinction to create topic models to illustrate different topic attention change of different sectors and companies. These differences will then be used to project stock returns.

4.1 Topic modeling using K-means

As a sanity check, we visualized the term frequency vectors generated from the previous section using PCA, in hopes to see that the semantic aspect of the annual reports will show some distinction.

We have selected data from four sectors that we think are distinct in nature: Energy, Material, IT and Finance. It is worth noting though, theses sectors are high level categories that may include different business. For example, Energy concerns business surrounding oil, gas as consumable fuels as well as equipment and service providers. Finance includes financial services, consumer finances, insurance and estate and mortgages. IT sector includes telecommunication, semiconductors and software providers.

Figure 8 below shows the term frequency vector PCA projection. As we can see the sectors are quite cleanly separated in different clusters, with Energy, IT and Finance standing out. and Material overlapping in between Energy and IT. With the added features, we could see that words related to the topics stand out. The Energy cluster has words such as gas, oil and drilled; the IT sector shows software; the Fiance sector has deposit, estate and mortgage. This further proves that term vectors does capture the different sectors.

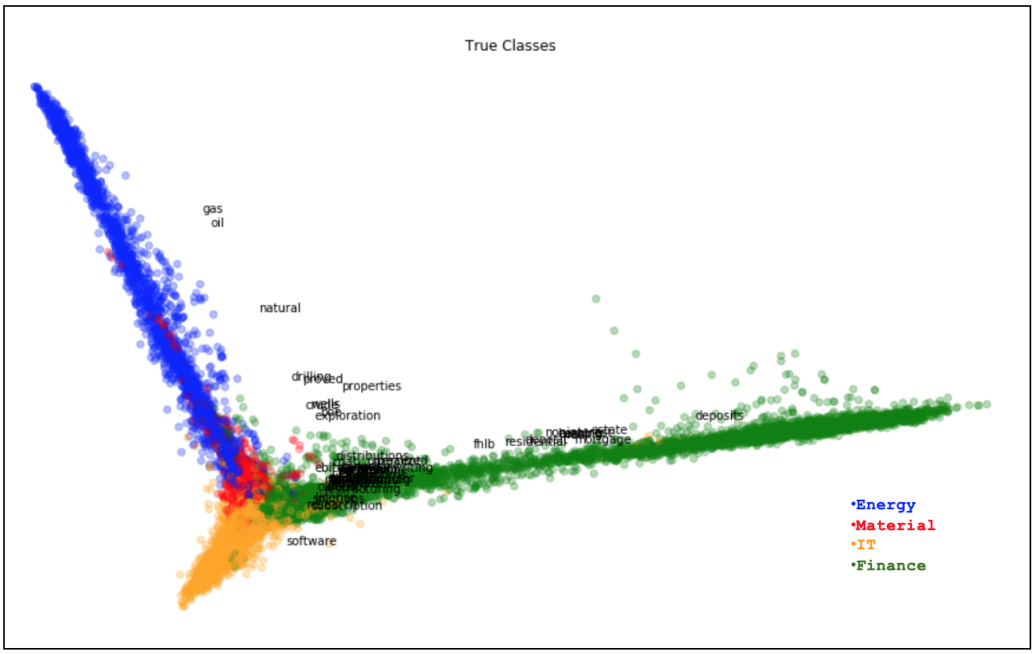

Figure 8: PCA visualization of TF vectors of selected sectors corpora

We then move on to cluster the annual report corpora with the selected sectors using K-means. Figure 9 below shows K-means clustering using 4 topics and 3 topics.

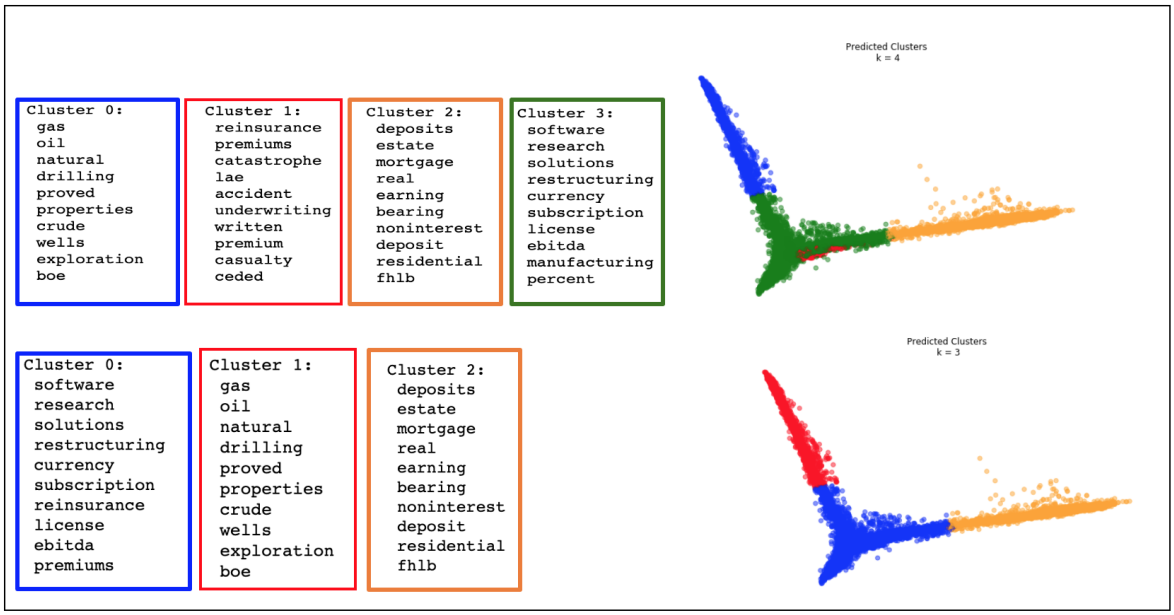

Figure 9: PCA visualization with predicted cluster of selected sectors corpora

The homogeneity score of k=4 is 0.345, and 0.298 for k=3. As we can see from above, K-means did not predict the document category very well. The clusters are somewhat different in the overlapping area. We can get a sense of why this is from the term vector PCA projection. The true clusters are almost like separate axes joined together with some overlap; k-means might not be the best algorithm to capture the centroids of clusters lying this way. Rather, the clusters are more like three groups forming a triangle in the 2D PCA. Despite the partial separation of the clusters, we see a good sense of words associated with each topic. In both cases of k=3 and k=4, we see gas, oil natural and drilling standing out as one cluster for the energy sector, while finance is well captured by deposit, estate and mortgage. The part that k-means may find difficult is to separate the overlapping IT and Material. In k=4, we don't see such crisp separation in that cluster 1 and 3 which supposedly represent Material and IT have somewhat questionable words. Granted, words like software and research is standing out; words like solution, license manufacturing and accident, casualty, premium could be for either sector. In k=3, we unsurprisingly see IT and Material essentially merging as a cluster with cluster 0 having words from cluster 1 and 3 from k = 4. This further shows k-mean's inability to separate the two.

4.2 Topic Modeling using LDA

After exploring the corpus using k-means, we then turn our attention to LDA to see if the separation of sectors might emerge. To achieve this goal, we conducted 4 experiments to better LDA's performance.

- Experiment 1: LDA using raw frequency count bag of words with the selected sectors

For this experiment we first built an LDA model based on 4 sectors, anticipating that the distinct semantic features captured in the Tf-idf cloud and term vector PCA will make sectors showing up as topics possible. However, to our dismay, we do not see human decipherable topics of words. We then turn our attention to the sub-sectors as topics, suspecting that because the corpus are all annual reports, they might have latent topics such as company profit, company operation, company vision ... etc. lurking in the background of the corpus. These aspects shares similar rhetoric, i.e. uses similar words and might even allows for monochromatic separation for the topics. The specific words for LDA with topic numbers equal to 4 and 8 are shown below in Figure 10 and Figure 11:

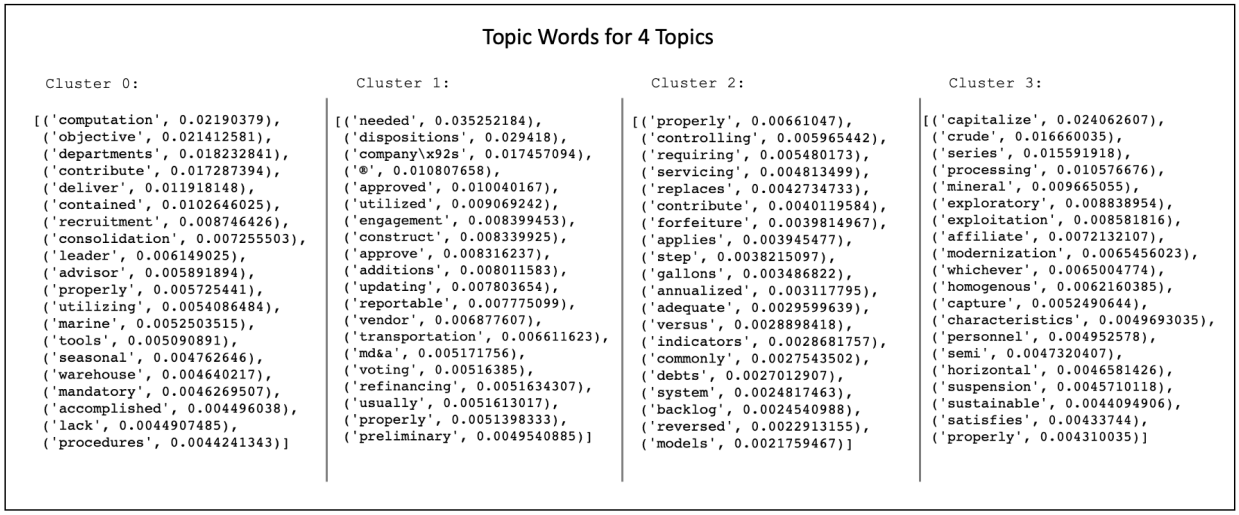

Figure 10: Topic words for 4-topic LDA with raw word frequency count

We can see that 8-topic LDA did not show sectors as topics as well. Rather, it contains more fine-grained clusters similar to the 4-topic LDA, further confirming the LDA is picking up another set of topics even though we set the number of topics align with the number of sectors and the number of sub-sectors. Take the

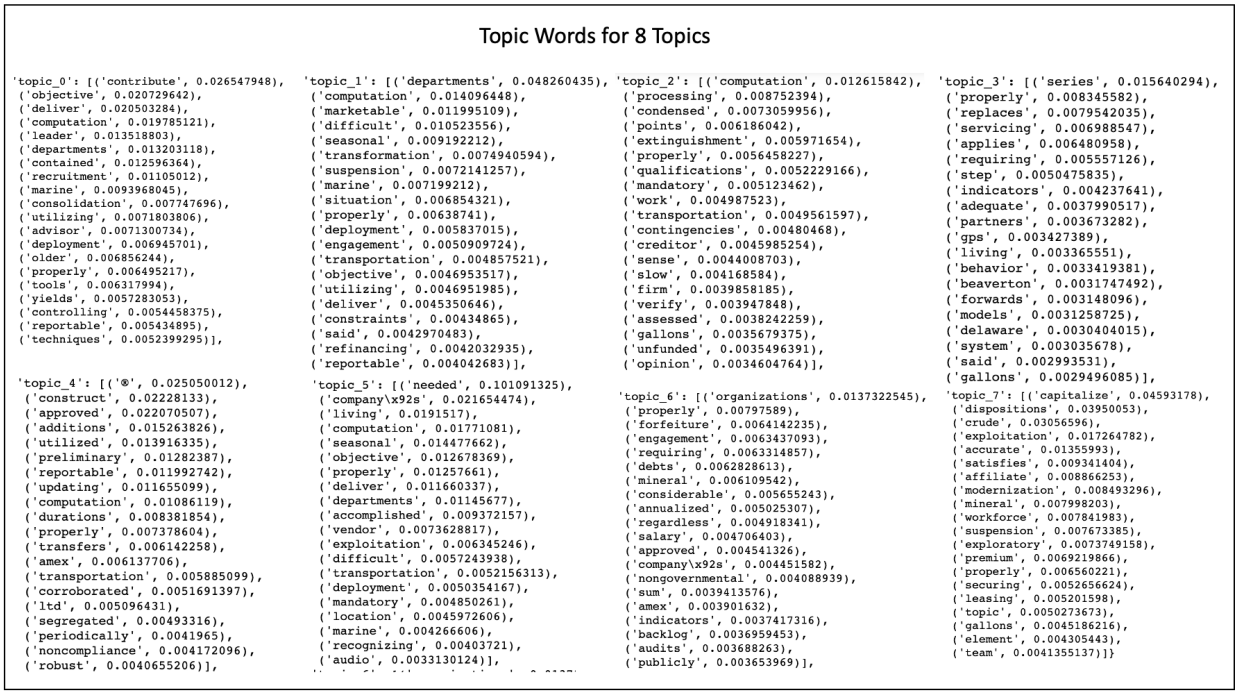

Figure 11: Topic words for 8-topic LDA with raw word frequency count

4-topic LDA as an example, we can see cluster 0 contains words that are related to company leadership, with words like objective, departments and leadership. Cluster 3 may have something to do with the work of the companies, with words like exploratory, capitalize, processing, mineral and modernization.

We also plotted topic attention across time to look for each sectors to see if there exists some pattern (Figure 12 and Figure 13). The change in topic attention from the 4-topic LDA and 8-topic LDA is somewhat stable within each sector, showing consistency over time. This is imaginable because it is sector-based. Topic attention between each companies are aggregated and the variances are canceled out. This does not serve our purpose of predicting stock return very well, but does confirm that the sectors have different semantic content showing in different topic dimensions picked up by the 2 LDA models.

To better serve the purpose of stock return prediction, we checked the company based topic attention change. Figure 14 shows the topic attention change across every issue date of the reports. We do see drastic variance in topic attention from the example IT company, which will be more meaningful to use for stock return prediction. For now, we continue on our quest for separating the sectors.

Figure 12: Topic attention by year for each sector from the 4-topic LDA. The topic attention change are aggregated sector-based results, which shows consistencies across time and distinct patterns for different sector, but does not serve our purpose of stock return prediction.

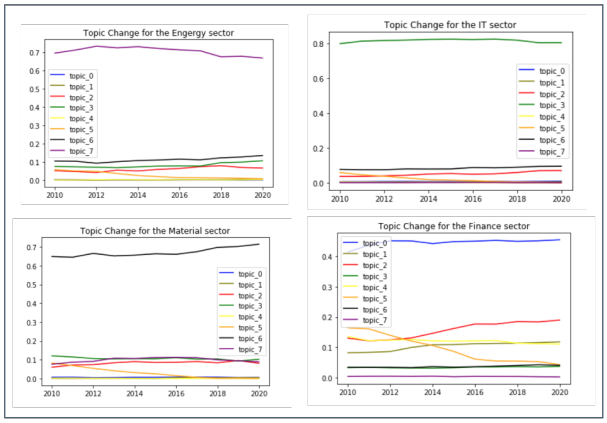

Figure 13: Topic attention by year for each sector from the 8-topic LDA.

- Experiment 2: LDA using bag of words of tokens whose IDF is above threshold

As the the LDA model with raw word count did not capture the sectors or decipherable topics, we decided to utilize the TF-idf counts to preprocess our data. Our intuition is to reduce the dimension of the multinomial distributions in the LDA model by selecting unusual words (Figure 15). The entire corpus then, would only contain words that are "unusual enough", excluding any annual report specific "stop-words". It is worth noting though, we only used the IDF scores as a selection condition. To build the model, we refered the selected words back to the original bag of words to the retain mathematical logic of LDA. Our model with pre-selected words based solely on IDF, however, did not gave meaningful

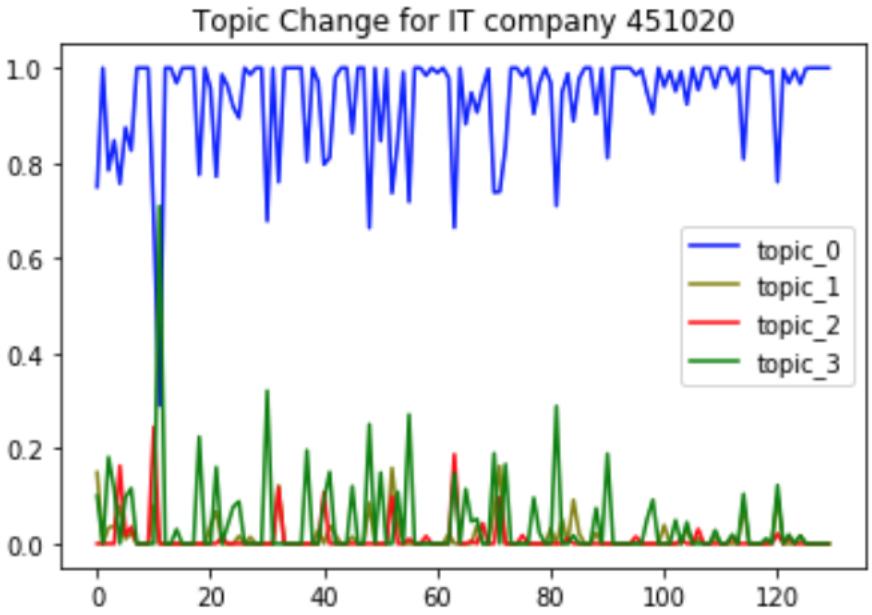

Figure 14: Topic attention change by issue date for IT company 451020 from 4-topic LDA

results.This experiment is taken further in the next step, experiment 3.

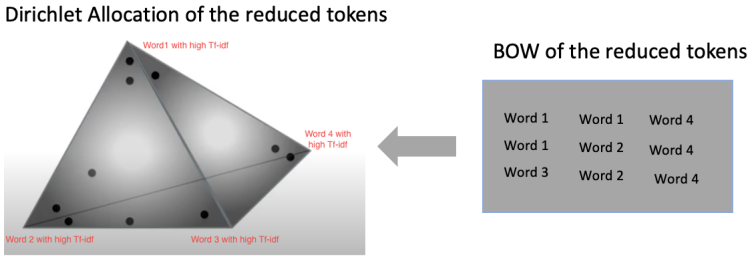

Figure 15: The Dirichlet allocation and multinomial distribution of tokens in the LDA model reduced in dimension

- Experiment 3: LDA using bag of words of tokens whose TF-IDF is above threshold for each of the document

Since our previous experiment did not render exciting results, we went further and processed the corpus on the document level. We reduced the size of each documents by only keeping the tokens of Tf-idf above mean. This approach saves training time, but is also reduces information of the corpus. This is a trade-off we are willing to take as it reduces the size of the vocabulary while increasing

the representativeness of the words in relation to the document, which can then exaggerate the uniqueness of each documents. After shifting through the tokens in each documents, we referred back to the raw frequency count so as to not compromise the idea of LDA.

Contrary to our previous assumption, reducing the dimension of the LDA model still did not show topics according to the high Tf-idf words. Exerperiment 3 rendered topic words highly similar to experiment 2. The 3-topic LDA model with pre-selected Tf-idf token shows the latent topics of document functionality. Figure 16, shown below, is the results of the LDA model that shows most decipherable topics to the humans. As we can see topic 1 shows all the symbols, number and dates. Topic 2 shows words related to connectivity: bridge, expand, user, exhibition, while topic 3 contains words like suffer, tax and testing.

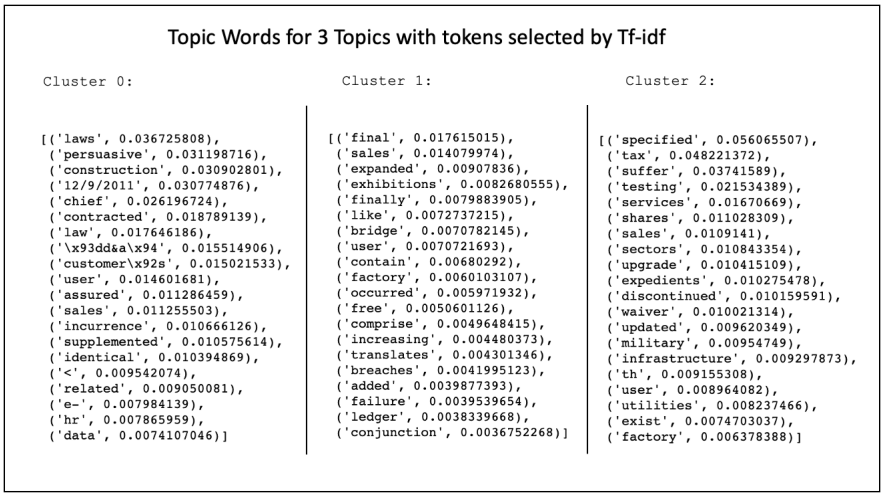

Figure 16: 3-topic LDA using selected token BOW based on tokens' Tf-idf values within each document

- Experiment 4: LDA using Tf-idfs values directly on the entire corpus

Although we have been avoiding feeding the LDA model with Tf-idf values, at this point, we are extremely tempted to do so because we want to explore the possibility of using LDA to pick up sector topics and not just latent topics. Building the LDA model with Tf-idfs on 10 topics on the entire corpus gave us reasonable separation based on some sectors, while some other topics remain undecipherable to humans. We thought this is a good combination of topics because we have topics representing the sectors which human can easily decipher, as well as topics

showing the latent feature of the annual report, which could be of good use in predicting stock return.

From the results shown in Figure 17, we see clustering of energy with natural, gas, oil and drilling, of telecommunication with comcast, television and satellite, of medical sector with trial, research, candidate, clinical, of material with gold, silver, ounce and mining and so on. We can also see some functional clusters, with symbol, fiscal, inventory and currency under a topic, and with advisory, mutual, inventory under another. Although undecipherable, these functional latent topics could serve our purpose of predicting stock return.

We did a sanity check to see if the topic attention change for this model also captures enough variance for prediction. As shown below in Figure 18, we see a lot of variance in topic attention captured for this single IT company throughout the time.

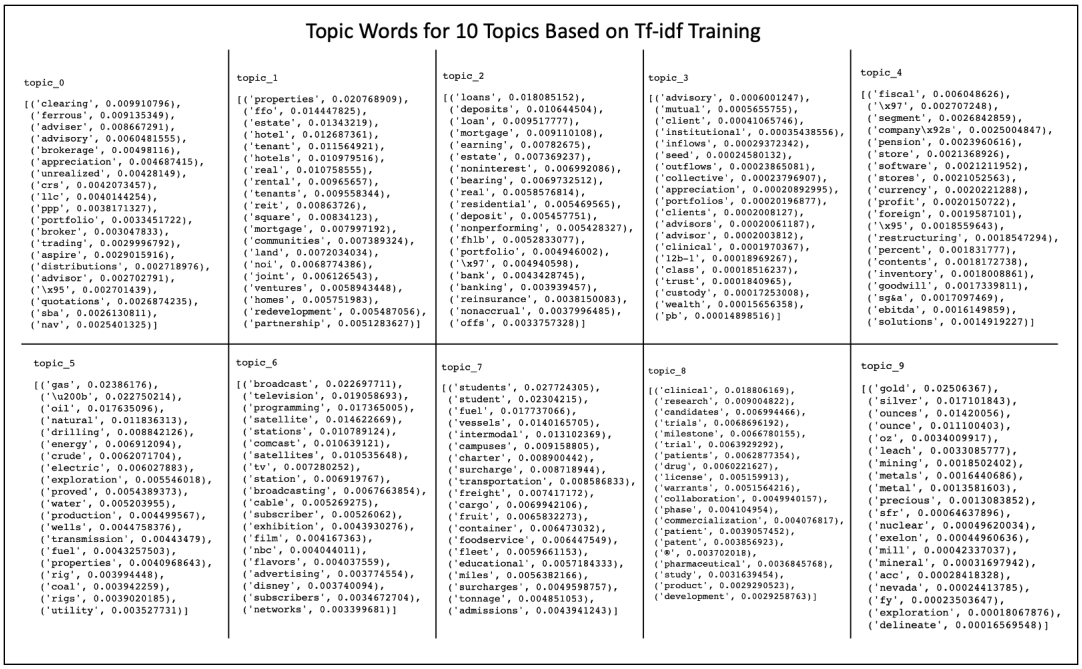

Figure 17: 10-topic LDA using Tf-idf value for each token within each document

4.3 Vectorizing topic attention for stock return prediction

In a similar fashion to section 4 of this report, we build a logistic regression model for stock return prediction using the LDA model described in experiment

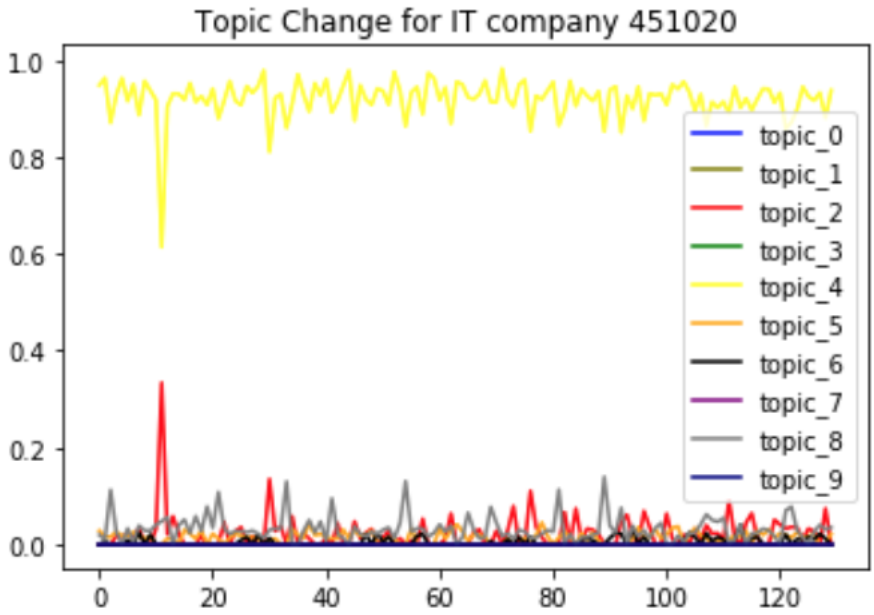

Figure 18: Topic attention change by time for a single IT company from the 10-topic LDA using Tf-idf values

4 as it captures company level variance, is trained on the entire corpus, and has both decipherable topic and undecipherable latent feature of each document. Our prediction model gave prediction accuracy of 0.549, slightly better than using the Tf-idf vectors, although still far from ideal. This further highlights the fact that predicting stock return is an extremely complex task; there are far more factors into play, but we feel enlightened to see semantic features may have some part in it and has the potential to play a bigger role with the appropriate reprocessing and modeling.

5 Word Embedding Analysis

We are also interested in the word embedding vector spaces of the four industries (Energy, Materials, Financial, Information Technologies (IT)) separately. Our goal is to capture some clustering feature by the visualization of the embedding space. Specifically, We adopted Google's word2vec algorithms that best describes corpus words in their local linguistic contexts, and mapped the trained vectors into a 3-D space using unsupervised clustering algorithms for visualization.

5.1 Data Pre-processing and Word2Vec Model Training

We followed the previous sections and used the [name] data, and used the `lucem_illud` for sentence tokenization and normalization. We adopted the default Word2Vec setting in the *gensim* library, and train the model using `gensim.models.word2vec.Word2Vec` command. The training output vectors are of 100 dimensions. After the training process, we also filtered words before the visualization. To make the lexicon we explored more representative, we excluded words that appeared in less than 20 articles, and words that appeared in more than 60% of the articles. (The articles mentioned here refer to the entire corpus covering the four industries.) After such a filtering strategy, there were about 4,000 words left in the text for each industry.

5.2 Clustering and Vector Space Visualization

For visualization, we used the *TensorFlow Embedding Projector* [1] that graphically represents high dimensional embeddings. [2]

We first looked at the default PCA clustering provided by the TensorFlow projectors. For all four industries, there were no specially patterns observed. The result is shown in Figure 19

We further applied t-Distributed Stochastic Neighbor Embedding (t-SNE), a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. Considering the the average dataset size is about 4k for each industries, we tuned the t-SNE with 'perplexity' [3] 35, learning rate 1, and didn't introduce any supervision. We trained for 900 iterations for each industry, and observed obvious convergence in the visulization.

The t-SNE visualization result is shown in Figure 20. For the energy, materials, and IT industries, we find that the t-SNE results give us a clear "mushroom-

[1] `https://projector.tensorflow.org/`

[2] The data for the projector are saved in `https://github.com/ZhouXing19/ContAnalysisFinalProj/tree/master/tsvs`. Please download the data to your local directory, go to `https://projector.tensorflow.org/`, and upload the metadata and vectors on the left panel on the webpage.

[3] "A second feature of t-SNE is a tuneable parameter, "perplexity," which says (loosely) how to balance attention between local and global aspects of your data. The parameter is, in a sense, a guess about the number of close neighbors each point has." – *How to Use t-SNE Effectively, https://distill.pub/2016/misread-tsne/*

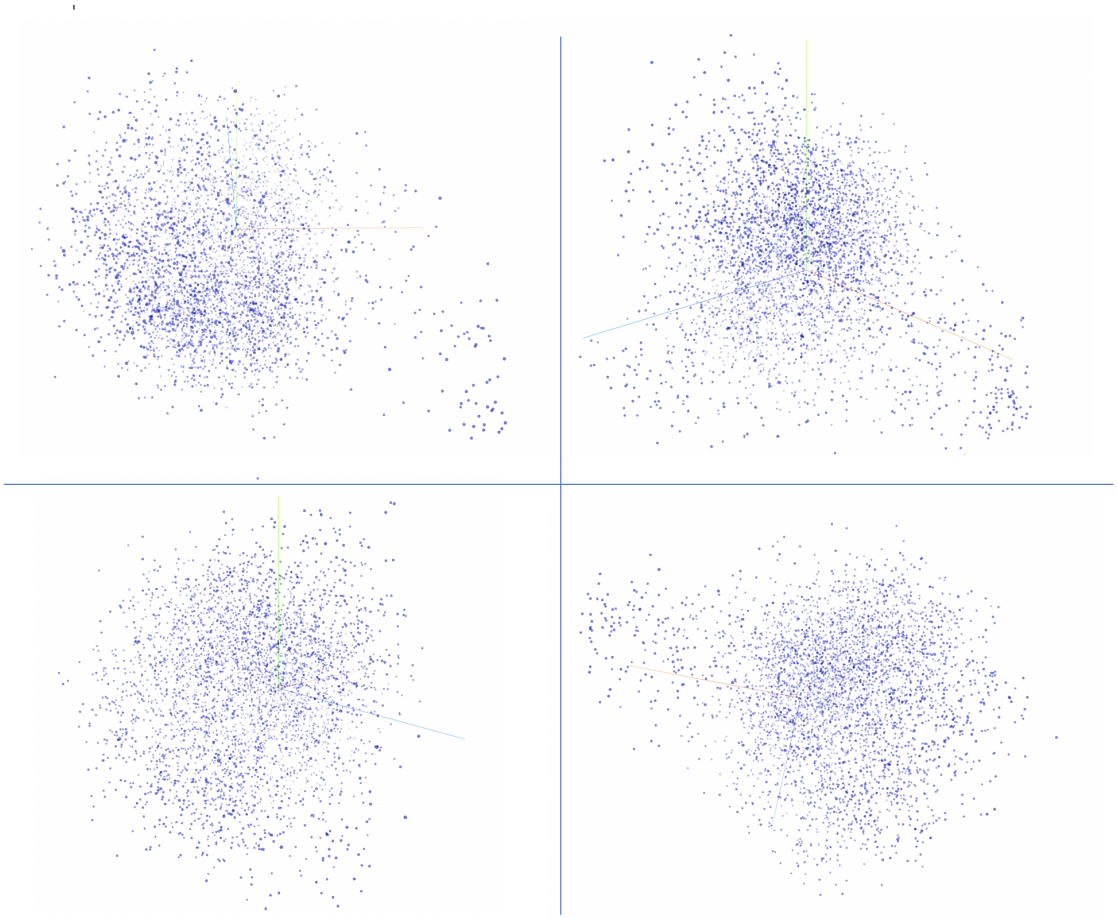

Figure 19: PCA Clustering for Four Industries: Finance(Upper Left), IT(Upper Right), Energy(Lower Left), and Materials(Lower Right)

like" structure: a small group of words are clustered in areas far away from the majority of words. While after isolating the small cluster from each industry of these three, we couldn't conclude the semantic commonalities of these words.

For the financials industries, we observe two small clusters that are distant from the majority of words. One clusters (top left in fig []) are abstract words share little common ground, while the other cluster are all words of American cities. We suspect this is because there are sentences in the original text of the financial industry that mention the cities in which the company's subsidiaries or outlets are located, and these sentences tend to list the names of these cities consecutively, so we see that they form their own separate little cluster.

Figure 20: T-SNE clustering (left) and isolation of words in the distinct small cluster (right) for Energy, IT and Materials (Top to bottom).

5.3 Dimension Projection

We can also project vectors of some key words to an arbitray semantic dimension. We specifically looked at two dimensions: 1. The dimension with risk against uncertainty; 2. The dimension with financial profit against social responsibilities.

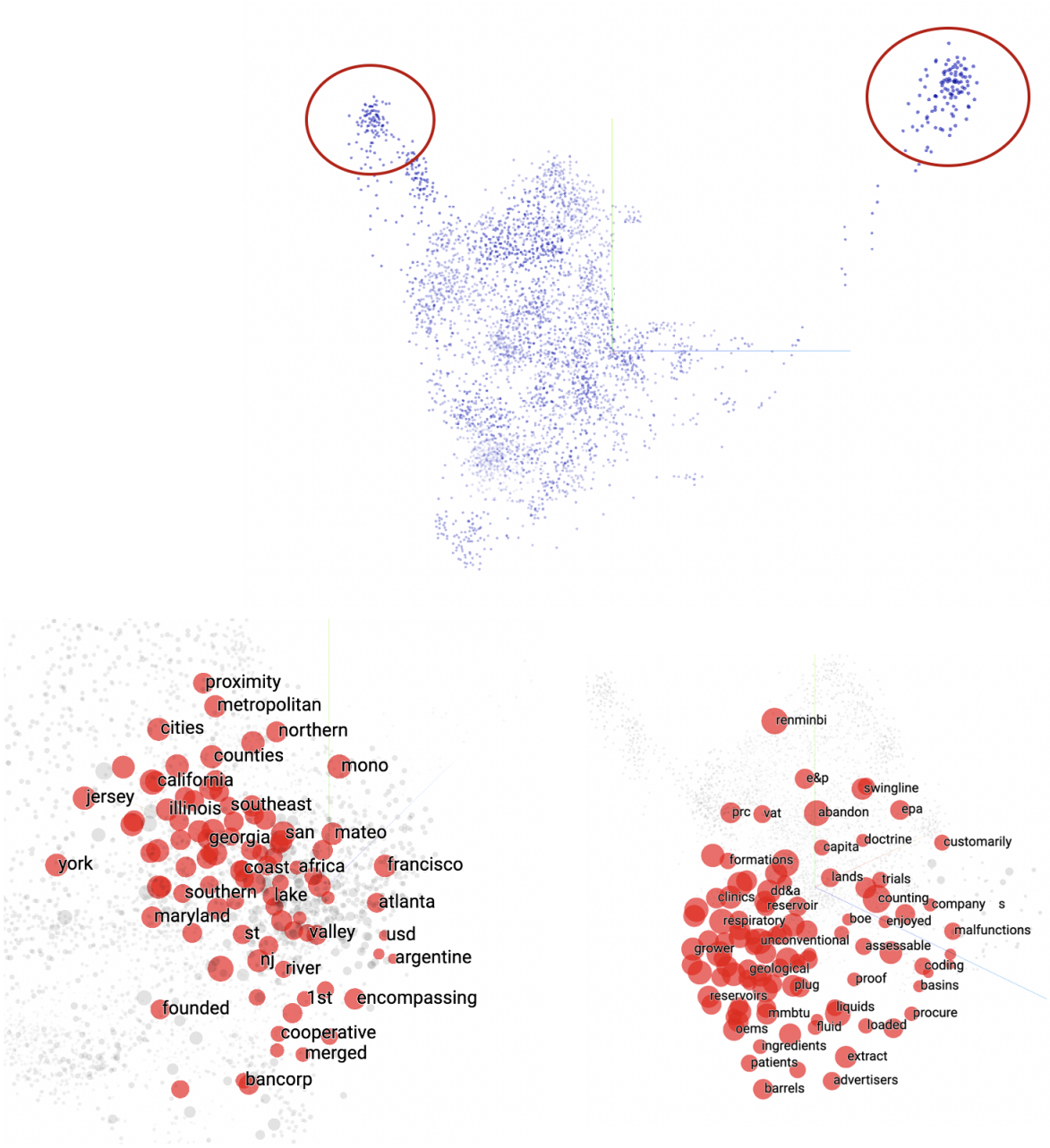

Figure 21: T-SNE clustering (left) and isolation of words in the distinct small cluster (right) for Finance.

5.3.1 Risk - Uncertainty Dimension

We followed the idea about the distinction of "Rish" and "Uncertainty" stated by Frank Knight, an idiosyncratic economist, from his 1921 book, *Risk, Uncertainty, and Profit.*[4] As Knight saw it, an ever-changing world brings new opportunities for businesses to make profits, but also means we have imperfect knowledge of

future events. Therefore, according to Knight, risk applies to situations where we do not know the outcome of a given situation, but can accurately measure the odds. Uncertainty, on the other hand, applies to situations where we cannot know all the information we need in order to set accurate odds in the first place.

"*There is a fundamental distinction between the reward for taking a known risk and that for assuming a risk whose value itself is not known,*" Knight wrote. A known risk is "*easily converted into an effective certainty," while "true uncertainty," as Knight called it, is "not susceptible to measurement.*" . [4]

Inspired by the Knight Uncertainty topic, we built up a dimension with words directly related to "Risk" (`[['risk', 'risks', 'risky', 'risking']]`), and words directly related to "Uncertainty" (`['uncertainty', 'uncertainties', 'uncertain']`), as the two ends.

We also looked into the distribution of the following key words in the Risk-Uncertainty dimension: `['political', 'regulatory', 'financial', 'interest', 'rate', 'country', 'social', 'environmental', 'operational', 'management', 'legal', 'competition', 'economic', 'compliance', 'security','fraud', 'operational', 'operation', 'competition']`. We selected words with consideration to cover a wide range of topics, including politics, financial conditions, and common topics of risk in corporation management.

We projected these key words from each word2vec embedding Figure 22.

We found that the overall distribution of words in the finance industry is more neutral and far from both poles, while the number of words in the remaining three industries is similarly distributed in this dimension. For example, the word "competition" is considered by all four industries to represent more of a risk, i.e., a quantifiable impact, while the word "environment" is considered by all four industries to be more of an unknown distribution. "uncertainty". For most of the terms, their distribution varies considerably across industries. We are particularly interested in the term "political": for the IT and materials industries, the term has a relatively strong association with quantifiable risk; for the energy industry, however, political-related topics always seem to be tightly linked to unquantifiable uncertainty. The bifurcation of the word politics across industries is also consistent with our intuition: for the Internet and technology industries, in recent years, political-related topics have always been associated with negative news: for ex-

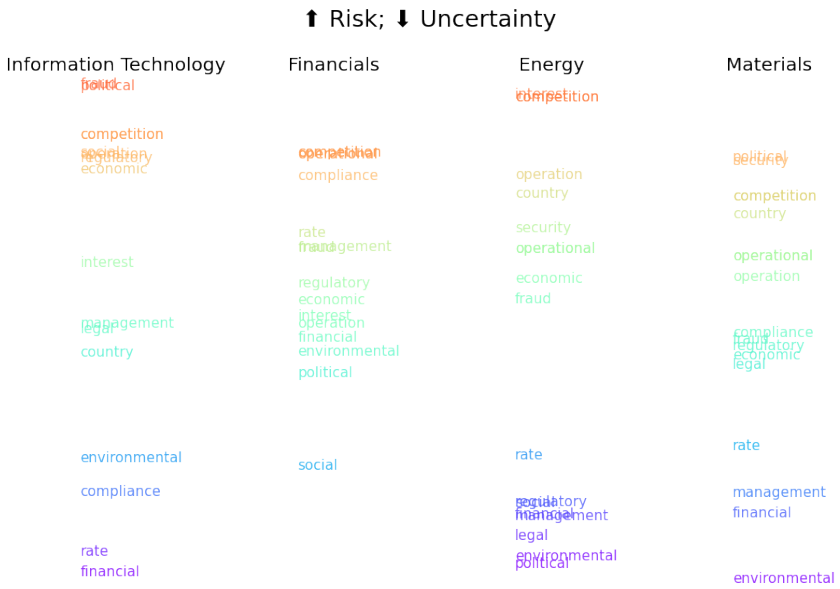

Figure 22: Projections of key words to the Risk-Uncertainty dimension.

ample, the rise of social media in social campaigns, especially elections, has raised concerns about its influence; U.S. technology giants face EU information controls; some countries have banned some technology companies for political reasons services for their citizens for political reasons, and so on. For the energy industry, however, there is also a strong connection to political events, but in many cases, the industry has benefited greatly from political events.

5.3.2 Financial Profit - Social Responsibility Dimension

We also weigh the trade-off between business interests and corporate social responsibility for companies in different industries. Therefore, we define the profit-social responsibility dimension and define two levels of this spectrum using related terms. Specifically, we use the terms `["revenue", "income", "profit", "growth", "boom", "interest"]` to target business profit and `["social", "environment", "governance", "responsibility", "diversity"]` to label social responsibility. We also pay extra attention to the following keywords: `["employment", "employee", "employees", "product", "products", "people", "management", "operation", "system", "goods", "goal", "vision", "value","customers", "clients"]`. These keywords cover the topics of business management, product

development, employee relations, and social responsibility. Similar to the previous subsection, we project the trained Embedding space of each industry onto the profit-social responsibility dimension. Our results are shown in the Figure 23.

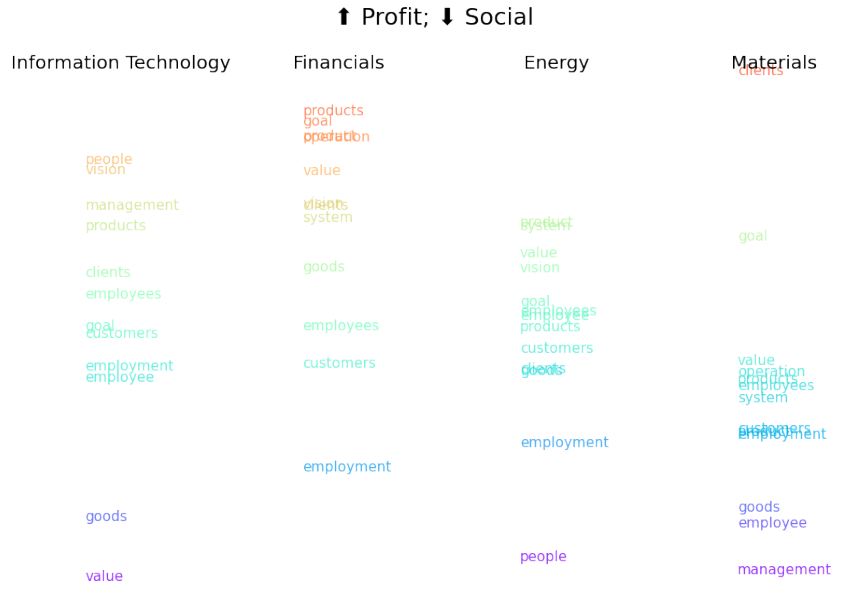

Figure 23: Projections of key words to the Risk-Uncertainty dimension.

We found that, overall, keywords were significantly more strongly associated with business profits in the model for the finance industry than in the other three industries. In addition, when looking at each term in detail, it is interesting to note that some of the terms also have a very clear polarized distribution across industries. For example, the word "people" is very close to the business profit end of the spectrum in the IT and finance industries, but in the energy industry, it is close to the social responsibility end of the spectrum. We guess this is because many products in the IT and finance industries are related to social networks, and "people" is the main target of these products, so they are mostly related to business growth.

6 Conclusion

We leveraged the Management Discussion and Analysis (MDA) section of the 10-K filing to conduct text content analysis from around 40,000 documents for three

main tasks: 1. to characterize four industries (energy, finance, IT, and materials) and predict stock returns for each market based on word frequency counts; 2. to perform topic modeling for these four industries; and 3. to use word2vec model for word embedding analysis. Although the predictive performance of this corpus is not significant, the dataset is informative in topic modeling and industry classification with LDA models. In addition, word2vec models trained in each industry corpus gave different results in terms of word embedding dimension projection. In future work, we will dig deeper into the potential of the MDA text dataset in stock return prediction.

References

[1] Lin William Cong, Tengyuan Liang, and Xiao Zhang. Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information. *Interpretable, and Data-driven Approach to Analyzing Unstructured Information (September 1, 2019)*, 2019.

[2] Fuwei Jiang, Joshua Lee, Xiumin Martin, and Guofu Zhou. Manager sentiment and stock returns. *Journal of Financial Economics*, 132(1):126–149, 2019.

[3] Zheng Tracy Ke, Bryan T Kelly, and Dacheng Xiu. Predicting returns with text data. Technical report, National Bureau of Economic Research, 2019.

[4] Frank Knight. *Risk, Uncertainty and Profit*. Number 14 in Vernon Press Titles in Economics. Vernon Art and Science Inc, July 2013.

[5] Frank Hyneman Knight. *Risk, uncertainty and profit*, volume 31. Houghton Mifflin, 1921.

[6] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65, 2011.

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