Sub-Industry Risk Classifier based on Risk Sentences

Abstract

In today's intricate business world, corporations are faced with sector-specific risks, while investors confront even more uncertainties that may impact their investments. Conventional risk assessment categorizes sectors and sub-sectors, yet fails to account for nuanced risk variations among similar businesses. This paper introduces a fresh approach: sector risk classification through textual analysis. By employing natural language processing, we extract insights from diverse sources to build a holistic risk classifier, augmenting financial data with qualitative information. This hybrid model empowers better decision-making for corporations and investors, offering a nuanced understanding of sector risks in a dynamic landscape.

Introduction

We want to derive a systematic method to identify the sector a risk factor belongs to based on a given risk sentence in the 10K. The intended purpose is to more accurately identify specific terms or sentences that are unique to a particular sub-industry, which may be challenging for a human to identify without many years of experience. This will help risk managers and investors better understand the specific risks in an industry even when they may appear to be similar.

Our project uses data from the companies from four different GICS sub-industries. GICS is an industry taxonomy developed by MSCI to classify sectors[1], and is split into Sector, Industry Groups, Industries and sub-industries.We focus on companies in the GICS FInancials sections, more specifically the four sub-industries - Diversified Banks, Asset Managers & Custody, Consumer Finance & Investment Banking/Brokerage.

On the sub-industries level, companies often have very similar business models e.g. Both derive revenues from similar sources despite different classification by GICS. Despite this, each sub-sectors has different risk factors that are often looked over by investors who pool together all financial institutions as one category, and can lead to misjudgement during the investment process.

Especially at smaller funds where investors may only have the capacity to separate investments into the Sector level - so for example, an analyst at a fund may be assigned to look at Financial companies, there may be the tendency to develop a risk framework for the entire sector, but miss the idiosyncratic risk factors that exist in each sub-industry.

Background

We took inspiration from our research from the paper *Extraction and classification of risk-related sentences from securities reports*[2]. The model uses 5000 sentences from Japanese companies to develop risk assessment into 5 factors. The results of this analysis gives us the confidence that Natural Language Processing techniques are effective in classifying financial data, especially in the context of Risk Types.

The paper *Risk Factors that Matter: Textual Analysis of Risk Disclosures for Cross-Section of Returns* (2018)[3] also attempts to use 10Ks to identify companies Risk Factor disclosures into a predefined list of risk factors and to classify them accordingly. Especially in the modern era with proliferation of data, there is an ever increasing need for efficient information processing and there have been numerous studies to use companies' textual risk disclosures to classify risk factors.

There have been numerous studies conducted regarding risk factors, but very few that help investors discover nuances between different risk factors of industries that are very similarly classified. We feel this is important for investors as identifying hidden risks in a specific sub-industry can improve the overall risk management process.

Approach

Data

Our primary source of data comes from 10K annual reports of the SEC EDGAR[4] database. Form-10 K filings have been mandatory since the passage of the Securities Exchange Act of 1934, and is an important disclosure document for investors and the public, allowing them to make informed decisions about investing in publicly traded companies. We specifically look at the section 'Item 1A: Risk Factors'.

Using the sec-api, we scrapped the specific section 'Item 1A' which provides the data. We applied this to all the recent 3 year 10Ks of all the companies in the four categories *Asset Management & Custody Banks, Diversified Banks, Investment Banking & Brokerage and Consumer Finance*.

Lastly, we split the text into sentences separated by full stops and saved it to a csv file. The categories are given a label from 0 - 3. The companies included are:

- **Asset Management & Custody Banks:** Blackrock , BNY Mellon , Franklin Templeton, Invesco, Northern Trust, State Street
- **Diversified Banks:** JP Morgan Chase, US Bancorp, Wells Fargo, Citigroup, Bank of America

- **Investment Banking & Brokerage:** Goldman Sachs , Morgan Stanley, Charles Schwab Corporation, Raymond James
- **Consumer FInance -** American Express, Capital One, Discover Financial, Synchrony

We managed to extract 25,000 sentences from 3 years of 10K 1A Risk Factor data for each company. However, we extracted an additional two years from the companies within the sub-sector *Investment Banking & Brokerage* as the Risk Factor sections seemed to be shorter than other sub-categories.

The distribution of sentences that we will train on are:

Label	**Sub-category**	**Number of Sentences**
0	Asset Management & Custody Banks	6325
1	Diversified Banks	5582
2	Investment Banking & Brokerage	7020
3	Consumer Finance	6610

Baseline

The 10Ks are traditionally read by Investment professionals in the field, and as both group members have had experience working in major financial institutions, we decided to manually label a random sample of sentences. This is also a method used by *Huang & Li*'s(2010) paper, where the authors classified 25 risk factors based on text and concluded that human analysts' understanding often has higher accuracy than computers.

Both group members each labeled 100 sentences and compared to true labels. The average result of our performance was 66%, which we will use as our baseline.

Modeling

We ran different models to test which had the best accuracy for our task in hand.

The models that we tested were:

- BERT
- Logistic Regression
- SVM (Support Vector Machine)

- Random Forest
- CNN
- FinBERT

Results

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
BERT	0.81	0.87	0.87	0.87
Logistic Regression	0.74	0.75	0.75	0.75
Support Vector Machine (SVM)	0.74	0.74	0.74	0.74
Random Forest	0.95	0.95	0.95	0.95
CNN	**0.99**	**0.95**	**0.96**	**0.95**
RNN	0.75	0.76	0.75	0.75
R-CNN	0.66	0.70	0.66	0.68
FinBERT	0.68	0.82	0.81	0.81

Most of our models outperformed our baseline. While traditional machine learning algorithms like Logistic Regression and Support Vector Machine showed improvements over the baseline, Random Forest emerged as the strongest performer with an accuracy of 0.95, effectively predicting all classes. Its ability to capture complex non-linear relationships and handle feature importance contributed to its success.

However, given the intricate nature of financial language, we also explored BERT and FinBERT. However, FinBERT[6], a modified version of BERT that is trained on a corpus of financial data by Professors at HKUST, performed worse than BERT, and only managed to obtain an accuracy of 0.68 and F1 score of 0.81 (BERT Accuracy 0.81, F1_score 0.87). This is in contrast to the literature where FinBERT performed significantly better than BERT in the corpus it was trained on.

CNN scored highest in all metrics, which was slightly surprising to us, as it is a neural network that is not specially trained to classify text, let alone financial information. After running 10 epochs the top train accuracy was 0.9946 and also had a high validation accuracy of 0.96. This is in contrast to the BERT models which had higher validation accuracies compared to trains.

RNN outperformed the baseline but R-CNN obtained similar results. Our RCNN model which was trained using GloVe 50B 6D performed worse than BERT. This could be because GloVe is based on static embeddings and does not capture context as effectively.

From the above, it seems that CNN would be most effective in correctly classifying, but Random Forest also performs well, and may be used as an alternative if computational resources are a concern.

Conclusion/Next Steps

An important conclusion that can be drawn from the project is that NLP based models can be used to help assess risk factors and inform decision making from text - be it 10Ks or articles abouts companies or industries. Natural Language Processing (NLP) plays a crucial role in assessing risk factors in various industry sectors by extracting valuable insights and patterns from unstructured textual data. NLP techniques enable the analysis of vast amounts of information from not just 10K but other sources too like news articles, social media, financial reports, and industry publications. By processing and interpreting this data, NLP can identify and monitor potential risks such as market trends, economic fluctuations, regulatory changes, and emerging threats. In the finance sector, for example, sentiment analysis can gauge public perception and predict market movements.

For next steps, we think the model can be further refined:

- Include more keywords into the analysis
- Extend beyond the finance sector into other industry sectors

References

[1] MSCI. “GLOBAL INDUSTRY CLASSIFICATION STANDARD (GICS®) METHODOLOGY”

[2] Fujii, M., Sakaji, H., Masuyama, S., & Sasaki, H. (2022). Extraction and classification of risk-related sentences from securities reports. International Journal of Information Management Data Insights, Volume 2, https://doi.org/10.1016/j.jjimei.2022.100096

[3] Lopez-Lira, Alejandro, Risk Factors That Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns (January 3, 2023). Jacobs Levy Equity Management Center for Quantitative Financial Research Paper, Available at SSRN: https://ssrn.com/abstract=3313663 or http://dx.doi.org/10.2139/ssrn.3313663

[4] U.S Securities and Exchange Commission , EDGAR-Search and Access https://www.sec.gov/edgar/search-and-access

[5] Li, Feng, Textual Analysis of Corporate Disclosures: A Survey of the Literature (February 7, 2011). Journal of Accounting Literature, Forthcoming, Available at SSRN: https://ssrn.com/abstract=1756926

[6] Huang, Allen and Wang, Hui and Yang, Yi, FinBERT - A Large Language Model for Extracting Information from Financial Text (July 28, 2020). Contemporary Accounting Research, Forthcoming, Available at SSRN: https://ssrn.com/abstract=3910214 or http://dx.doi.org/10.2139/ssrn.3910214

[7] L. Li, L. Xiao, N. Wang, G. Yang and J. Zhang, “Text classification method based on convolution neural network”,2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, 2017, pp. 1985-1989.

[8] A. Hassan and A. Mahmood, "Efficient Deep Learning Model for Text Classification Based on Recurrent and Convolutional Layers," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, 2017, pp. 1108-1113.

[9] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. “GloVe: Global Vectors for Word Representation”.