What do Firms Disclose in their ESG Report? Evidence from Topic Modeling and FinBERT*

Tan Tingshuang†

April 17, 2023

Abstract

This project aims to uncover the contents disclosed in companies' ESG reports by analyzing the ESG reports published by firms listed in HKEX, using tetxual analysis techniques including topic modeling (LDA and BERTopic) and FinBERT. The project wishfully helps stakeholders to understand firms' ESG report better from textual analysis angle and provides some insights on ESG reporting related standard setting.

Keywords: Environmental, social and governance (ESG), disclosure, NLP, topic modeling

1 Introduction

In recent years, there is a growing desire by many investors for sustainable investment. The size of the green investment market in US has grown significantly in the last decade from 22.9 trillion dollars in 2016 to 40 trillion dollars in 2020 (US SIF, 2020). Along with the desire for green investment is the demand for information about firms' ESG activities. Because of this, many organizations around the globe have issued reporting standards for ESG activities which intended to imporve the ESG reporting practics (Christensen et al., 2021). In Hong Kong, HKEX firstly introduced the ESG Reporting Guide in 2013 and required manatory disclosure of ESG reports in 2015. Over the years, HKEX continues to enhance the disclosure requirement. For example, on 5 November 2021, HKEX issues Guidance on Climate Disclosures to encourage listed firms to disclose their Climate-related Financial Disclosures; on 10 December 2021, the updated version of ESG Reporting Guide is released, requiring firms to publish their ESG reports at the same day as annual reports (PwC, 2022).

*All codes, model and raw data can be found at `https://www.dropbox.com/sh/smf70ff1e0kh9of/AACsiJQcuiMK07HRNqtGknNqa?dl=0`

†UID: 3035533422. Email: ttshuang@connect.hku.hk

There is also a proliferative literatures in accounting and finance studying the effect of ESG reporting to different stackholders. Christensen et al. (2021) comprehensively documented related literatures about the effect of ESG reporting on firms' value, risks, and stock returns etc., to equity holders, debt holders and standard setters etc.

Leveraging NLP techniques including topic modeling and FinBERT ESG classification, this project intends to uncover the specific contents in firms' ESG reports for firms listed in HKEX. In addition to the traditional LDA topic modeling, BERTopic[1] modeling, an emerging technique that uses language embedding and class-based TF-IDF will also be applied to the analysis. Further, the project will use the ESG classifier from FinBERT, a language model based on BERT and mainly tackling NLP tasks in the financial domain [2], to examine the proportion of three dimensions (i.e. environmental, social and governance) in an ESG report. Hopefully after the analysis, we can obtain a more comprehensive and accurate understanding about firms' ESG reports.

Section 2 will brief the data collection process using web-scraping and how the corpus is constructed. Section 3 will cover the results from LDA and BERTopic. Section 4 report the results from FinBERT ESG classification. Section 5 concludes and documents some limitations.

2 ESG Report Corpus Construction

2.1 Web-scraping

Companies listed in HKEX periodically disclose their reports or news on HKEXnews[3]. Using python packages including selenium, bs4, requests etc., we can obtain information (including ESG report name, release time, report url etc.) of firms' ESG reports by inputing file types, keywords, and time range [4]. A glance of the scraped report information is shown in Figure 1. The pdf files can then be downloaded using the scraped urls. In total, 6421 PDFs which are about 40G have been downloaded from the HKEXnews.

The next task is to parse the pdfs to txt files based on which we can create the ESG report corpus. For most companies listed in HKEX, their disclosures are reported in both English and Chinese. This project only investigated into the English texts[5]. Hence, during the parsing process, some preliminary pre-process have been made: non-English (Chinese words) words are deleted; all digits are deleted; and texts are striped to prevent extra space. We lose 287 files during the parsing process due to formatting issues. In total, we get 6171 txts file[6] after the parsing process.

[1] See Grootendorst (2022) or `https://maartengr.github.io/BERTopic/index.html`

[2] See Huang, et al (2022) or `https://finbert.ai`

[3] `https://www1.hkexnews.hk/search/titlesearch.xhtml?lang=en`

[4] For details, please refer to the webscraping_report_information.py.

[5] Further research can also consider to analyze the Chinese version and compare the information revealed by different language. For example, Lang, et al. (2022) analyzed the differential treatments and local information advantage from the translation difference of the annual reports.

[6] The size is about 1G and the raw data are stored in the raw_txt.zip

2.2 Industry Information

Because the contents disclosed in the ESG reports are usually industry-specific (Christensen et al., 2021), we want to perform our analysis at industry level. It is somewhat annoying that I can't find any direct industry data about firms listed in HKEX after searching on official website. Instead, I can only find a chart at Hong Kong Economic Times[7] and aggregate the data with the Hang Seng Industry Classification System[8] (which is the official classification system used by HKEX) to obtain the firm-industry data.

After preparing those data, we can construct the ESG report corpus by join the report information data, raw txt data and firm-industry data (see Figure 2). The corpus construction process is summarized in Figure 3.

2.3 Descriptive Data

The number of ESG reports by fisical year is depicted in Figure 4. Since HKEX firstly introduced their ESG Reporting Guide in 2013 and mandated ESG disclosure in 2015, we can see the number of reporting witnesses a sharp increase in 2015 and keep increasing over the next few years. There are only limited observation in 2022, whcih is because, untill the time I scraped the reports (04/13/23), many firms have not disclosed their 2022 ESG reports yet. The number of ESG reports by different industry is displayed in Table 1.

3 Topic Modeling

The results from LDA and BERTopic on the constructed corpus will be presented in this section. More details are available in the Jupyter Notebook.

3.1 LDA Topic modeling

The corpus is processed through standard pre-processing, tokenization and lemmatization. I also include bigram (which occurs greater than 20 times in a document) to increase the intepretability. Further, I filter out tokens that occur less than 20 documents, or more than 50% of the documents. We can have a preliminary look at the corpus to see the heterogenous contents discolsed by firms from different industries. For instance, common tokens for CGN Power (01816) from enegy industry are "nuclear_power", "enegy", "hazardous_waste" etc, while on the other hand, for HEC Pharm (01558) from Healthcare industry have more tokens like "drug_administration", "generic_drug", and "intellectual_property" etc.

[7] `https://invest.hket.com/markets/industry`
[8] `https://www.hsi.com.hk/static/uploads/contents/en/dl_centre/brochures/B_HSICSe.pdf`

Due to the long running time, I give up performing the cross validation about choosing the right number of topics, which is one of limitations of this project. The analysis is performed using full sample with $K = 20$ and 30.

When $K = 20$, the average U_mass topic coherenece is -1.05. The WordClouds of each topic are generated and are shownin Figure 5[9]. Overall, the model performs well in terms of separability and interpretability as for each topic depicted, we can intuitively assign it a name from the keywords. We now take a closer look at some appealing topics:

Topic#0 Keywords: "pharmaceutical", "drug", "school", "university", "teaching", "exhaust_gas", "innovative" etc. This topic is about the pharmaceutical companies. It's surprising to see words like "school" "university", and "teaching" are also in the same topic. My guess is that the pharmaceutical companies might have close connection with university labs and may have collaboration with the professors as well. They also ca re about how to deal with exhaust gas during the pharmaceutical processes.

Topic#5 Keywords: "real", "estate", "law_republic", "urban", "smart", "public_welfare", "home" etc. This topic is related to firms in real estate industry. From the keywords, we can infer that they may like to cite the relevant law ("law_republic") and to stress their efforts in promoting public welfare.

Topic#9 Keywords: "bank", "poverty", "alleviation", "poverty_alleviation", "rural", "finance" etc. This reveals that firms in financial industry like to emphasize their efforts in poverty allivation in their ESG disclosure (which complies with the Social dimension in ESG reporting standard). As Chinese governement emphasized a lot about poverty alleviation especially in rural area plus bank as financial intermediaries plays important roles in poverty reduction, it makes sense that banks disclose more about this in their ESG reports.

Topic#12 Keywords: "engineering", "technological", "labor", "ecological", "scientific, "law_republic" etc. This topic is related to tech firms. "Labor" is among the top words in this topic, which is reasonable as these years, tech firms are facing controversy on their employees' overtime issue. They may want to justify in their ESG report by disclosing their efforts in complying with the labor law.

Topic#18 Keywords: "food", "retail", "hotel", "wellness", "plastic", "executive", "scheme", "ordiance" etc. This topic reveals the contents disclosed by service companies including catering and hotelling. Their focuses may include the use of plastic bag, customer's wellness, food safty ordiance, and executive governance as shown by the keywords.

[9]The order does not represent any ranking.

The above topics shows the disclosure practice within an industry. There are also more general topics from the generated topics that all firms concern about, such as carbon footprints, renewable energy, ecological economy, employee training, code conduct, and COVID 19 pandemic.

The results by setting $K = 30$ are quite similar. It has poorer coherence score compared with the case $K = 20$. I also performed LDA analysis by different years to see if the disclosed contents have evolution over time. However, the result is quite homogenous to full sample. If you are interested, the detailed results can be obtained in the Jupyter Notebook.

3.2 BERTopic

Compared with LDA, BERTopic is much easier and faster to run and it only requires simple preprocessing (which is excluding stopwords). We only need to fine tune the some parameters (like minimum topic size, which is the minimum number of documents in the resulted topics) to train the models. When setiing minimum topic size as 10, 100 topics are generated in our corpus (see Figure 6). The top topics are quite similar to what we obtained from LDA. Moreover, we can draw the hierarchical clustering of the model to vividly see the similarity between each topics. In the Jupyter Notebook, there are more results with different parameters.

I also tried the model in different industries including finance (see Figure 7), real estate (see Figure 8), and IT (see Figure 9), from which we can get more detailed information about each industries.

4 FinBERT

The ESG classifer from FinBERT can classify a sentence into four categories: "Environment", "Social", "Governance", and "None"[10]. It also assigns corresponding possibilty score to the classification. For example, for sentence "In June, the upgrade of water recycling pipelines was completed", the output from classification is {'label': 'Environmental', 'score': 0.96}, which means that with 96% probability, the sentence is about environment issue. Using the sentence tokenizer from nltk, we can then assign each tokenized sentence a category. Here, we assume the classification is accurate thus dismiss the information about the probability. By doing this, we are able to get the number of sentences in each category and further calculate the proportion for an ESG report. Although our data are at firm level, in this project, we will simply look at the average statistics by industry and years.

Figure 10 displays the average proportion of each category within each industry and Figure 11 is the average proportion of each category in each year. From the two charts, we can see different

[10] An intuition about the proportion of category "None" is that it could be a proxy for irrelevant information or noise in the reports.

dimension that different industries emphasize in ESG reports and the change in the proportion overtime.

Overall speaking, companies put most of their pen and ink on social aspects followed by environment issue. Governance ranks the lowest and sometimes are enven lower than the None_proportion. By industry, compared with others, companies in financial industry have more social and governance contents and less environment contents in their ESG reports. Their proportion for None category is also the highest, indicating that financial institutions may write more about ESG irrelevant information. On the contrary, firms in utility, materials and energy industry disclose more about envrionmental issue.

Over time, we can clearly see the proportion change. Environment and governance contents are increaseing while social contents have decreaseing trend. It is noticeable that the irrelavent contens (represented by None_proportion) are also declining over time, which might be a sign of increase in ESG disclosure quality.

5 Conclusion and Limitation

This project leverages the advanced NLP techniques to analyze the contents in the firms' ESG reports. By webscraping ESG reports from HKEX, a large ESG reports corpus is constructed. Topic modeling reveals many hidden information in the ESG reports, and FinBERT ESG classification managed to discover the distribution of the three dimensions in different industry and its change over time.

Despite fairly good results, there are still some parts that can be further improved. First, the Chinese characters are dismissed in our analysis. Secondly, in the LDA modeling, CV or other model choice methods could be applied to choose the best K. Lastly, in the FinBERT classification, nltk sentence tokenizer is used. However, by assessing the tokenized outcomes, I find SpaCy should perform better than nltk, but it will take much longer time to run. Due to the time limit, I give up on that.

Hopefully, this project can help stackholders to understand the current ESG reporting practice more comprehensively and accurately.

References

[1] Christensen, H. B., Hail, L., & Leuz, C. (2021). Mandatory CSR and sustainability reporting: Economic analysis and literature review. Review of Accounting Studies, 26(3), 1176–1248. https://doi.org/10.1007/s11142-021-09609-5

[2] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure (arXiv:2203.05794). arXiv. http://arxiv.org/abs/2203.05794

[3] Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text*. Contemporary Accounting Research, 1911-3846.12832. https://doi.org/10.1111/1911-3846.12832

[4] Lang, T., Stice-Lawrence, L., Wong, Y. T. F., & Wong, T. J., (2022) Differential Treatment and Local Information Advantage: Revelations from Translation Differences. USC Marshall School of Business Research Paper Sponsored by iORB, No. Forthcoming, http://dx.doi.org/10.2139/ssrn.3956105

[5] PricewaterhouseCoopers (2022). ESG Reporting Study for Hong Kong Listed Companies 2022. https://www.pwccn.com/en/issues-based/esg-report-2022.pdf

[6] US SIF. (2020). Report on US sustainable and impact investing trends 2020. https://www.ussif.org/files/trends%20report%202020%20executive%20summary.pdf

Tables and Figures

release_time	stock_code	stock_short_name	report_url	report_type
06/07/2016 16:32	1816	CGN POWER	https://www1.hkexnews.hk/listedco/listconews/s...	2015 Environmental, Social and Governance Repo...
05/07/2016 22:43	1558	HEC PHARM	https://www1.hkexnews.hk/listedco/listconews/s...	Environmental, Social and Governance Report 20...
30/06/2016 18:51	242	SHUN TAK HOLD	https://www1.hkexnews.hk/listedco/listconews/s...	SUSTAINABILITY REPORT 2015 (4019KB)
30/06/2016 16:55	3	HK & CHINA GAS	https://www1.hkexnews.hk/listedco/listconews/s...	Sustainability Report 2015 (6008KB)
23/06/2016 12:20	1185	CHINA ENERGINE	https://www1.hkexnews.hk/listedco/listconews/s...	CORPORATE SOCIAL RESPONSIBILITY REPORT 2015 (4...

Figure 1: Glance of Scraped Report Information

release_time	identifier	fiscal_year	stock_code	en_name	cn_name	ind_code1	ind_name1	ind_code2	ind_name2	ind_code3	ind_name3	esg_report
06/07/2016 16:32	2016_01816	2015	01816	CGN POWER	中廣核電力	40	公用事業	4000	公用事業	400010	電力	Stock Code: CGN Power Co., Ltd.*(A joint stock...
05/07/2016 22:43	2016_01558	2015	01558	HEC PHARM	東陽光藥	28	醫療保健	2810	藥品及生物科技	281010	藥品	YiChang HEC ChangJiang Pharmaceutical Co., Ltd...
30/06/2016 18:51	2016_00242	2015	00242	SHUN TAK HOLD	信德集團	80	綜合企業	8000	綜合企業	800010	綜合企業	The design concept is inspired by the traditio...
30/06/2016 16:55	2016_00003	2015	00003	HK & CHINA GAS	香港中華煤氣	40	公用事業	4000	公用事業	400020	燃氣	SUSTAINABILITYREPORT(Stock code:) INNOVATION ...
23/06/2016 12:20	2016_01185	2015	01185	CHINA ENERGINE	中國航天萬源	10	工業	1010	工業工程	101030	環保工程	CONTENTSAbout this Report Group Profile St...

Figure 2: ESG Report Corpus

Industry-level 1	Number of ESG reports
Consumer Discretionary	1533
Properties & Construction	1038
Industrials	810
Financials	548
Telecommunications	424
Healthcare	384
Materials	372
Consumer Staples	312
Utilities	292
Energy	223
Information Technology	53
Conglomerates	42

Table 1: Distribution by Industry

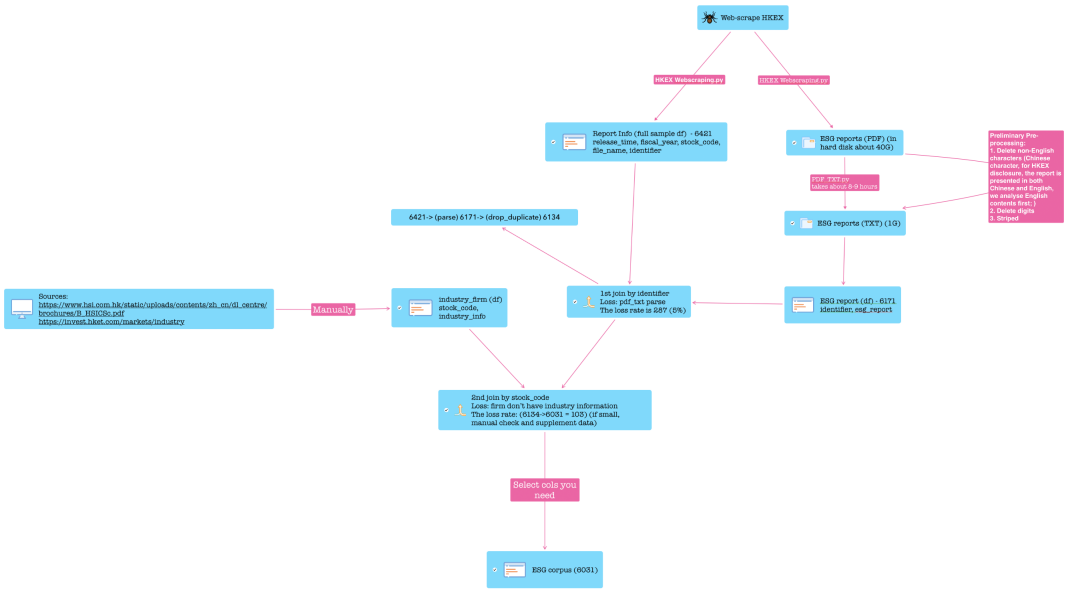


Figure 3: ESG Corpus Construction

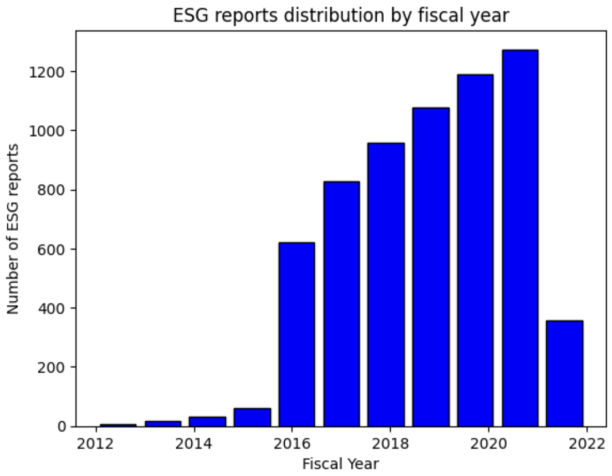


Figure 4: ESG reports distribution by fiscal year

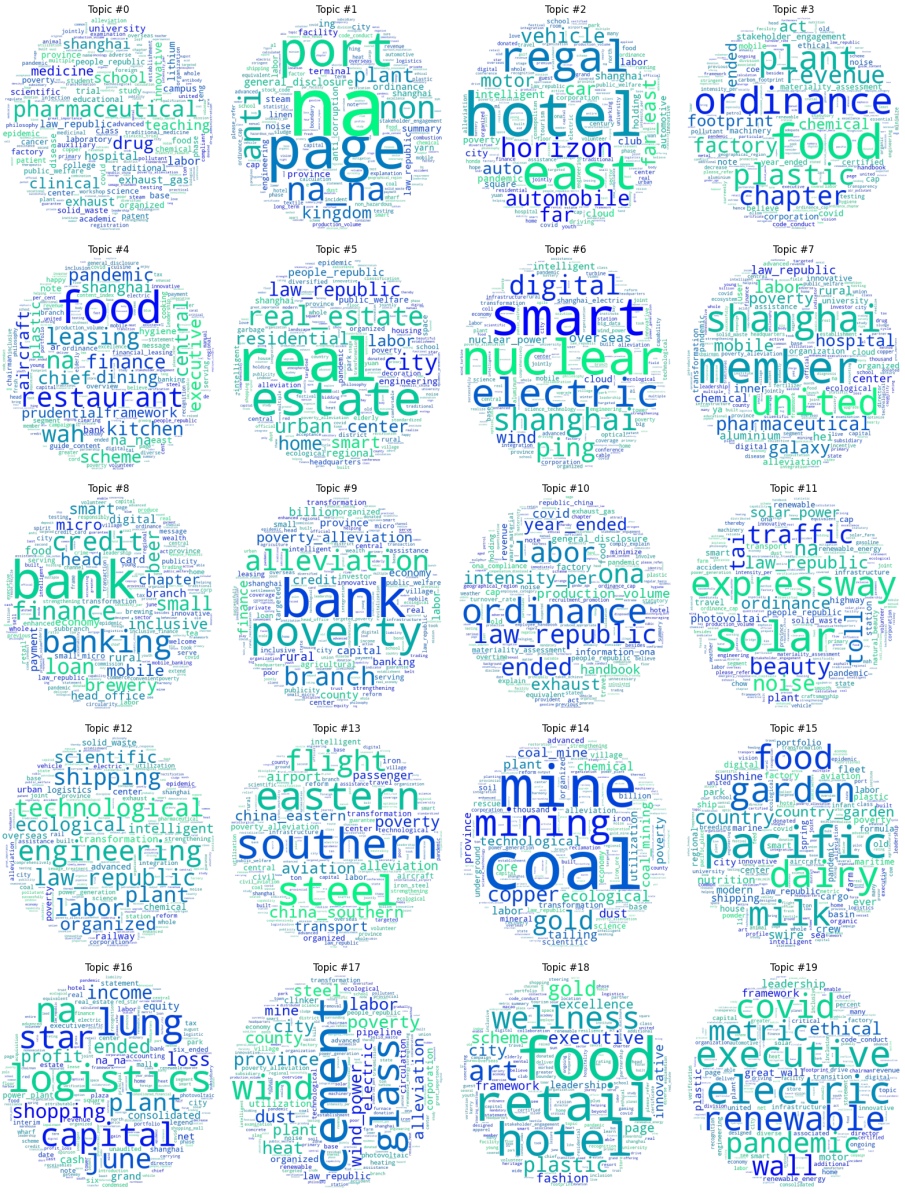
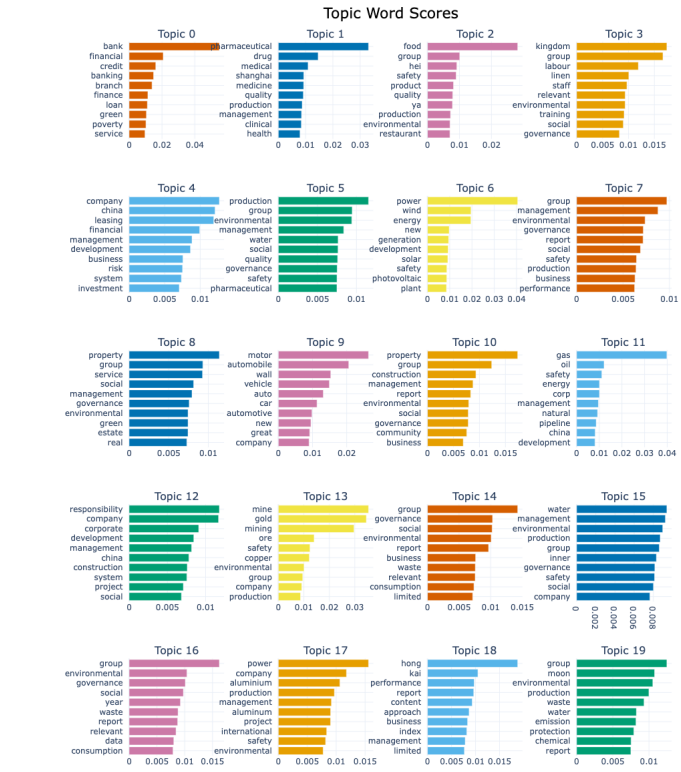


Figure 5: WordClouds for LDA TM ($K = 20$)



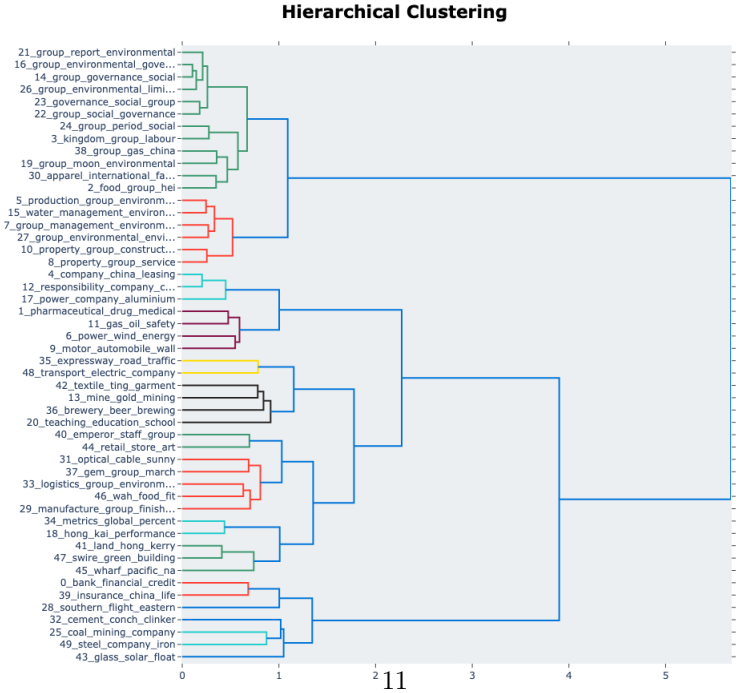
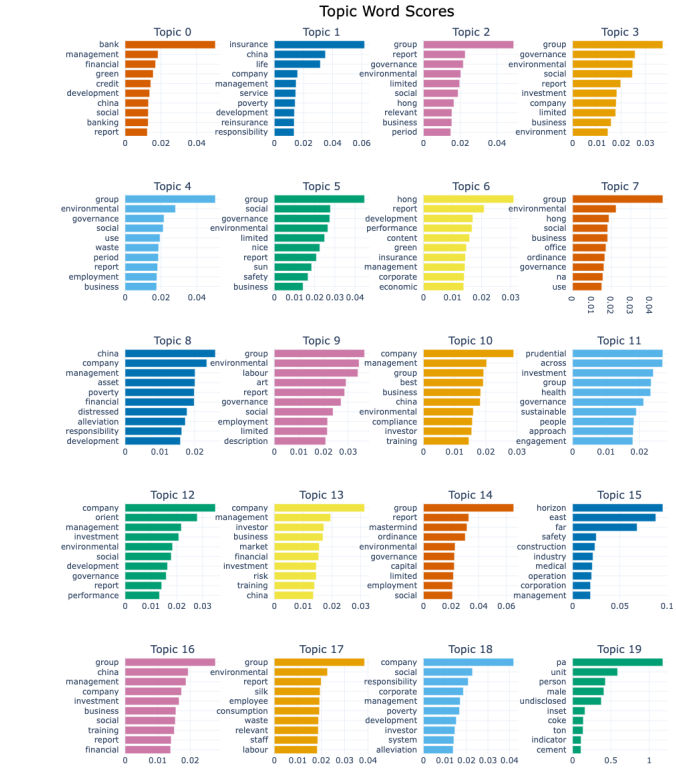


Figure 6: BERTopic (minimum topic size = 10)



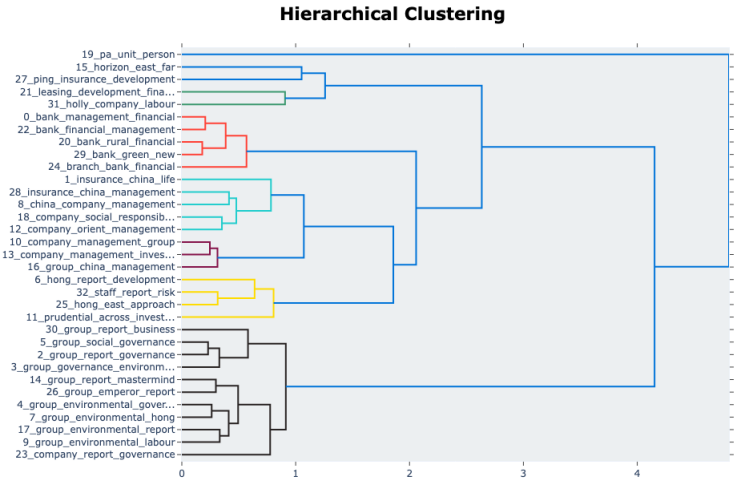
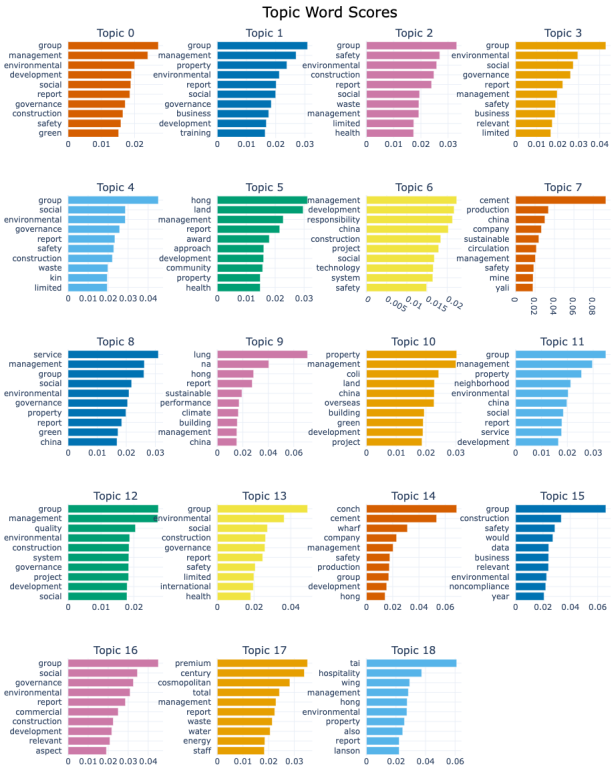


Figure 7: BERTopic Finance Industry



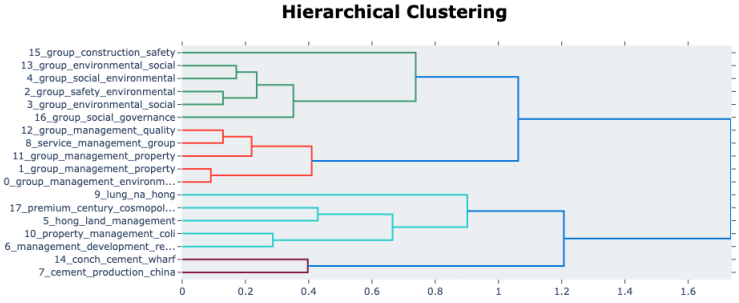
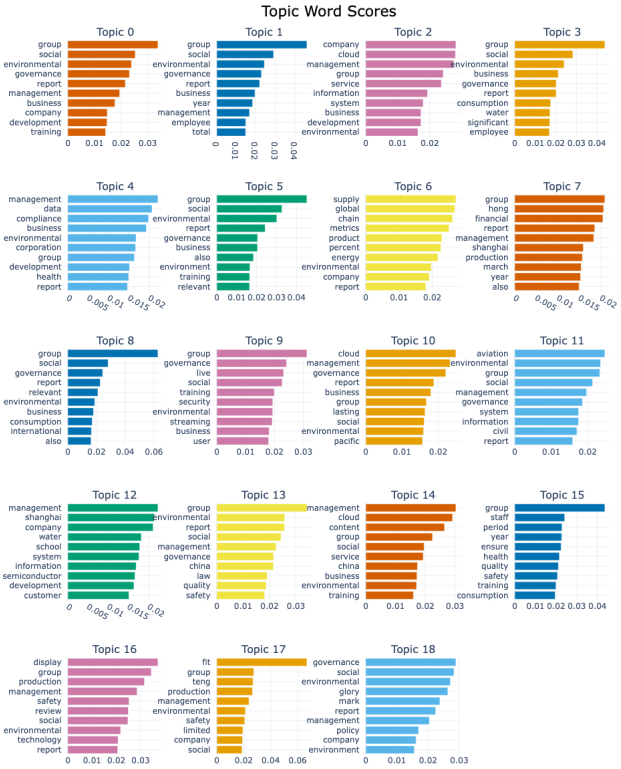


Figure 8: BERTopic Real Estate Industry



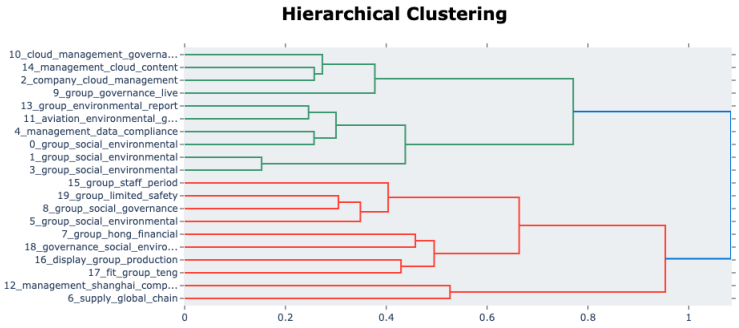


Figure 9: BERTopic IT Industry

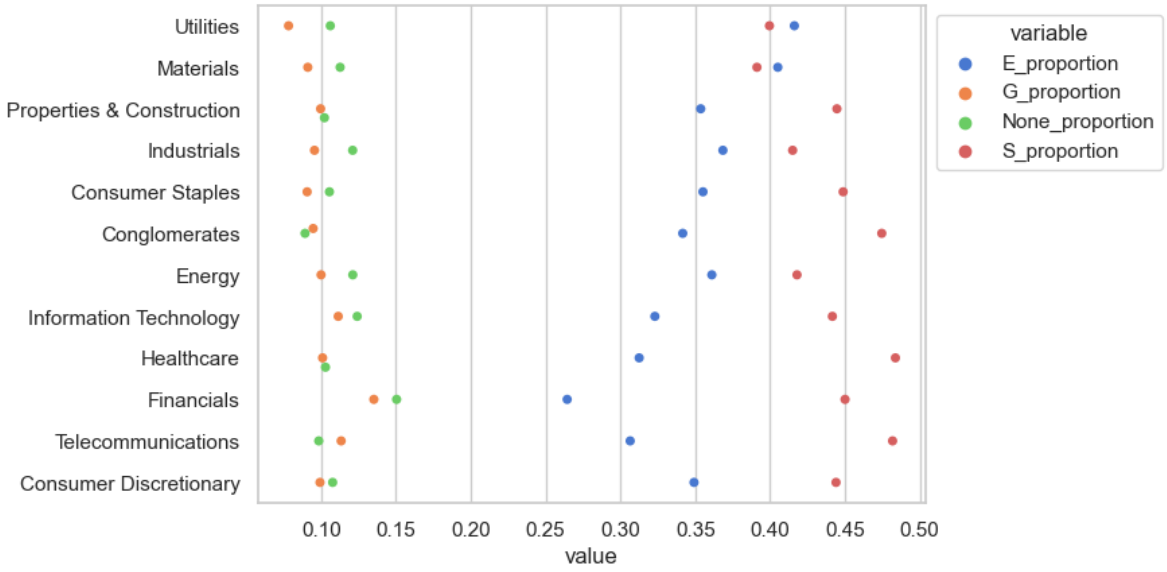


Figure 10: ESG Proportion by Industry

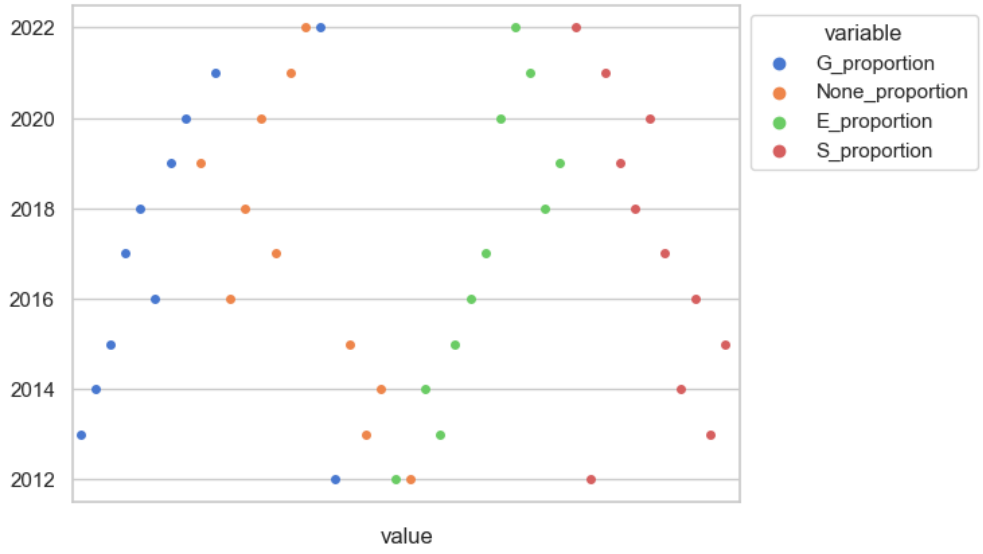


Figure 11: ESG Proportion by Year