Shining a Light on ESG with NLP

ABSTRACT

This report investigates the application of natural language processing (NLP) techniques to streamline the ESG (Environmental, Social, and Governance) scoring process by summarising annual sustainability reports. The study leveraged a dataset of 1200 ESG reports web-scraped from responsibilityreports.com. LDA was the primary model developed, with secondary research exploring how ESGBert can be used for feature engineering. Ultimately, the LDA model was optimised using coherence scoring and trained on 1200 reports, leading to a novel ESG scoring method based on the frequency of topic mentions. The results of this study were integrated into an interactive Streamlit app, which employs the trained LDA model to generate customised insights based on the user's uploaded report.

In summary, this report demonstrates the potential of NLP techniques to facilitate the ESG scoring process, improve report summarisation, and offer valuable insights for investors and stakeholders.

Source. https://www.azeusconvene.com/esg/articles/what-is-esg-issues-importance-and-initiatives

Henry Paremain

10.04.23

1. INTRODUCTION

1.1 Background

ESG (Environmental, Social and Governance) refers to a set of criteria used to evaluate the sustainability and ethical impact of a company or investment. ESG-related data is becoming increasingly more important, with investors demanding greater transparency on companies' activity in the ESG space, to help identify the risk of certain companies and allow for well-informed investment decisions. There are a number of agencies offering ESG scoring services, where companies are given a rating to reflect whether their activities in these areas are deemed to be favorable or unfavorable.

The current ESG rating system is flawed for a variety of reasons, including:

1. No standardisation - 8 rating providers (and growing), all giving different ratings.
2. Data on companies' ESG activity is hard to acquire.
3. The rating process is time-consuming and costly, requiring domain expertise.

1.2 Problem Statement

How can NLP techniques be used to summarise annual ESG reports and facilitate the ESG scoring process?

2. DATA COLLECTION

Relevant Notebooks:

- *1.1_text_collection*
- *1.2_score_collection*

2.1 ESG Reports

Publicly traded companies are required to publish an annual ESG report, summarising the key initiatives that the business has undertaken relating to the space. Various data sources are currently considered to score a company, one being the annual ESG report, which offers substantial insight into the company's operations and initiatives. A company with less to say about a particular ESG topic is suggestive of poor performance in the given area, whilst a company that is doing a lot of positive work will no doubt be reporting it.

Responsibilityreports.com is a website that hosts many of these annual reports. A web-scraper was created to gather annual reports for 1200 companies, which formed the body of data for this project.

2.2 Historical ESG Scores

A package called 'yesg' was used to obtain historical ESG scores from the YahooFinance API, which gets its scores from Sustainalytics. Initially, these scores were collected to see if this project could take a supervised learning approach and predict how a company's risk changes throughout the year. However, there was not enough data available, as ESG scores are only updated a couple of times a year. The historical scores which were collected were ultimately used as a comparison with the custom scoring created later on in the project.

3. DATA PROCESSING

Relevant Notebooks:

- *1.1_text_collection*

Having obtained the pdfs, the raw text needed to be extracted. Two packages were trialed - 'pyPDF2' and 'pdftotext'. The latter was preferred due to its ability to retain a similar paragraph formatting to the original pdf, so that when it came to splitting the reports, it was easier to identify natural breaks in the text.

When it came to separating the text, three main techniques were used. The first method looked at using the entire pdf as a single corpus of text, merging paragraphs into a single text. This was problematic when it came to topic modeling, as the model was unable to infer independent topics within the corpus. The second method aimed to split the report into sentences. This was suboptimal due to the messy nature of the raw text, with sentences often merging together. The third method aimed to split the report into paragraphs. This method produced the most coherent topics, and 'pdftotext' made the process of breaking up the paragraphs simple.

Having split the text into paragraphs, the text needed to be cleaned. Details of this can be found in Notebook 1.1, with a summary shown below in Figure 1.

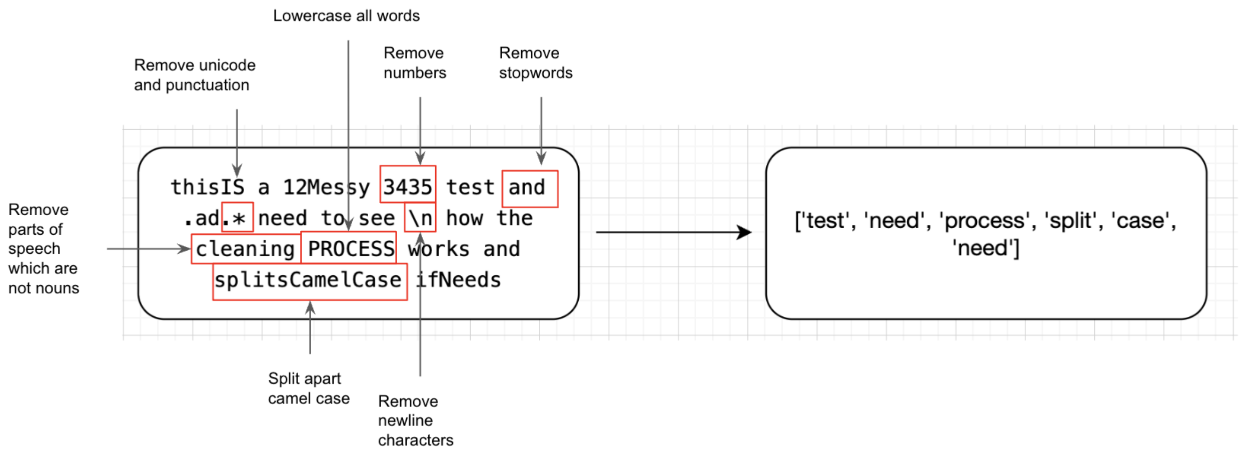
Lowercase all words
Remove unicode and punctuation
Remove numbers
Remove stopwords
Remove parts of speech which are not nouns
thisIS a 12Messy 3435 test and .ad.* need to see \n how the cleaning PROCESS works and splitsCamelCase ifNeeds
['test', 'need', 'process', 'split', 'case', 'need']
Split apart camel case
Remove newline characters

Figure 1 - Example to demonstrate preprocessing steps

4. EDA

Relevant Notebooks:

- *2.1_eda*

The primary aim of the EDA stage was to identify the type of words and themes contained in the bank of reports. The secondary aim was to optimise the text preprocessing. The charts which displayed keywords (such as Figure 2 below), gave a quick insight into whether the preprocessing was working and what needed to be refined. For example, some words such as 'environmental', 'sustainability' and 'report' kept appearing at the top of the most frequent list. These words provide no insight into the topics covered by a report, and so this meant these words (along with a few others) were ultimately removed from the corpus of text going forward.

Figure 2 is an example of the insights gained from grouping reports by industry and looking at the most common words. One can clearly see differences in the focus areas of each industry.

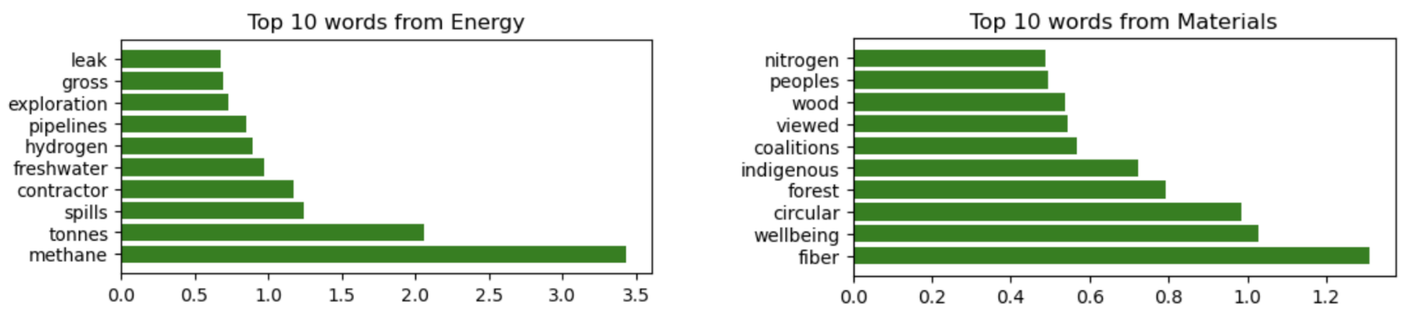

Figure 2 - Example insights from EDA

5. MODELLING

Relevant Notebooks:

- *3.1_lda_model*
- *3.2_esgbert_model*

Two different approaches were taken with modeling. The supervised learning approach explored clustering algorithms based on features produced by ESGBert. The second approach looked at unsupervised learning using LDA topic modeling.

Feature engineering utilised ESGBert, a language model pre-trained on a corpus of sustainability text. By feeding a report into ESGBert, one could gain the percentage of 26 different ESG metrics captured within the report. These percentages could then be used as features. KMeans, DBSCAN, and Hierarchical Clustering were carried out to explore if there were any obvious clusterings. KMeans proved to be the most useful, indicating 2 distinct types of company. One type of company focused much more on governance-related topics, whilst the other group focused on environmental topics. Although this was a useful insight, there was not enough data available to try to predict ESG scores and so this work was taken no further.

The LDA topic model was the second model to be explored. Various parameters were experimented with, including the impact of different preprocessing techniques (eg lemmatizing vs no lemmatizing); the difference between TF vs TFIDF vectorization; the impact of the number of topics; and how different dimensionality reduction algorithms impact the clarity of the topics. The tests concluded that the following setup led to optimal results:

- TF vectorization
- Lemmatization of tokens
- Only using nouns in the model
- Dimensionality Reduction = Principal Coordinate Analysis (PCOA)

When it came to optimising the LDA model further, Coherence Scores were used. Various papers report the link between a high coherence score and the interpretability of the words within a topic. This score varies depending on the number of topics chosen for the LDA model. Hence, to find the optimal number of topics, the Coherence Score was calculated for a range of topic numbers, with the optimal number being the model with the maximum coherence score.

The Relevance Metric of the pyLDAvis package was found to be a useful feature for improving interpretability. By setting this variable to 0, the words which were more specific to a certain topic were given greater weight and thus displayed higher up the pecking order. Hence, when it came to looking at keywords for a topic, one could gain a better idea of what the theme of the topic was.

Having identified the optimal setup for LDA, the model was trained on 1200 reports, split by paragraph. Each paragraph could then be fed into the model and a probability of the three topics (E S and G) belonging to that paragraph was output. For each report, the mean of these probabilities could be found for each category, to provide a 'score', representing the strength to which that particular topic was mentioned in the report.

6. CONCLUSION

The output of this project is summarised in an interactive dashboard built using Streamlit (run '4.1_streamlit_app.py'). The first page summarises key insights from the 1200 ESG reports from 2021. The second page allows the user to upload any ESG report and gain custom insights, including an ESG score; the most relevant paragraphs for each of the E, S and G topics; and the key topics identified by the model and the associated words. Both pages utilise the LDA model trained on 1200 reports.

The Streamlit app highlights the potential of the work summarised in this report. The ability to quickly gain an understanding of the relevant paragraphs of a long report has the potential to save a lot of time in the review process. This, combined with a 'score', allows a quick comparison between companies, to identify where companies are placing their attention.

7. FURTHER WORK

There are a number of routes that could be taken to help refine the results of this project, summarised below.

1. Network analysis to determine a company's most influential partners, based on mentions in news headlines. A weighted score could then be created to account for the impact of one company's ESG behavior on another.
2. Semantic analysis of social media mentions. By looking at tweets and other social media mentions of companies, one could account for whether they are mentioned positively or negatively in the context of ESG topics, and weight the score accordingly.
3. Further hyperparameter tuning of the LDA topic model.