Using Natural Language Processing to Analyze Financial Climate Disclosures

Anonymous Authors[1]

Abstract

According to U.S. financial legislation, companies traded on the stock market are obliged to regularly disclose risks and uncertainties that are likely to affect their operations or financial position. Since 2010, these disclosures must also include climate-related risk projections. These disclosures therefore present a large quantity of textual information on which we can apply NLP techniques in order to pinpoint the companies that divulge their climate risks and those that do not, the types of vulnerabilities that are disclosed, and to follow the evolution of these risks over time.

1. Introduction

In 2010, the U.S. Securities and Exchange Commission (SEC) published a guidance regarding the disclosure requirements that companies must comply with relating to the issue of climate change (U.S. Securities and Exchange Commission, 2010). This guidance stipulated that companies issuing stocks traded on U.S. markets were legally obliged to disclose any risks and uncertainties related to climate change, be it its physical impacts on assets or distribution chains, its legislative impacts on profit generation, the international accords affecting prices of raw materials, and even indirect market consequences such as supply and demand (Doran & Quinn, 2008).

While this is undoubtedly a step forward in transparency and accountability, the problem remains that SEC filings are hundreds of pages long and filled with financial and legislative jargon, which makes it hard to pinpoint the passages, if any, that address climate risk disclosures. In fact, a 2013 study found that of 3,895 U.S. public companies listed on major stock exchanges, only 27 percent mentioned climate change explicitly in their annual reports (Hirji, 2013). However, this study utilized keyword search, which leaves the door open for a more in-depth analysis using Natural Language Processing (NLP) in order to identify which companies are disclosing various types of climate risk and to what extent. In our study, we aim to focus on a subset of major U.S. companies in order to analyze the extent to which these companies disclose climate risk, what they divulge, and how this evolves before and after the 2010 ruling.

2. Methodology

We propose to bootstrap the process of analyzing climate disclosures by starting with a set of hand-picked keywords, then automatically expanding the vocabulary to be able to detect sections in the text that address climate risk. We then propose to extract information using state-of-the-art NLP tools from those sections focusing on targeted categories and concepts. This pipeline both streamlined and transparent to simplify human intervention for evaluation purposes and to guarantee the ownership of the analysis afterwards.

For our analysis, we will focus on 10-K reports, which are publicly available documents that are published on the SEC website [1] and that give a comprehensive summary of a company's financial performance on a yearly basis. While all companies traded on the stock market can (and eventually, will) be affected by climate change, we will start with companies from the S&P 500, the largest companies listed on the U.S. stock markets which represent about 80% of their total market value. We will focus more specifically on companies from sectors that are most obviously impacted by climate change: energy, real estate, oil and gas, utilities, and materials (Winn et al., 2011), for a total of 115 companies. We target the ten-year period between 2008 and 2018, to encompass a time frame before and since the SEC ruling. This will give us a total of 1150 documents to analyze.

2.1. Identifying Climate Disclosure Sections

Despite the SEC ruling, it is probable that many companies simply do not explicitly include climate risk disclosures in their annual findings. Therefore the first challenge of our work is finding the relevant sections within the disclosures that explicitly mention these risks. In order to identify these passages, we propose to start with a semi-supervised approach, constructing a climate dictionary from a small set of seed words selected manually, then automatically expanding these seed words to build a full-fledged vocabulary. The eight initial seed words that we propose in our approach are: *biodiversity, carbon, climate, ecology, environment, emission, pollution, sustainable*. Starting with this seed vocabulary, we propose to take each word and query a 300-dimensional word embedding model trained on Wikipedia

[1] https://www.sec.gov/edgar.shtml

and the Gigaword corpus (Pennington et al., 2014) to generate one vector per word. We use each vector to retrieve the ten closest words using cosine similarity, expanding this initial list of 8 seed words into 50 climate-related words (since many of the related words overlap), which we can use to identify relevant passages in the 10-K reports. The expanded list of words includes terms like: CO2, deforestation, ecological, greenhouse, pollutant, viability, etc. For a full list of vocabulary words, see the supplementary materials.

Subsequently, using our climate vocabulary, we propose calculate a *climate relevance score* for each section, which we can use in order to identify the sections that are likely to directly elaborate on climate-related risks, and rank the identified sections in each report by relevance. We propose that sections that contain one or more of the eight initial seed words be assigned a higher climate-relatedness score, whereas those containing words derived from the seed words rank slightly lower. We can then use this score to compare the reports of a single company within a given time frame to verify whether there was an evolution in the climate disclosures before and after the 2010 ruling, but also of companies within the same sector, such as energy or real estate, in order to measure different companies' degree of accountability with regards to climate risk.

2.2. Analyzing Climate-Related Discourse

Above and beyond identifying the sections of the 10-K reports that address climate-related risks, we also want to analyze what types of risks are being divulged by the companies. While there has been some research on analyzing climate disclosures, it has mostly involved keyword searches or manual analysis by experts (Gamble et al., 1995; Doran & Quinn, 2008). We propose to use more complex, statistical-based NLP methods such as named entity recognition, sentiment analysis, and information extraction.

As a first step, we propose to identify the types of entities that are being addressed in the climate disclosure sections. These entities can be both names of companies and actors, for instance British Petroleum, an oil and gas company, or Bayernoil, the name of one of its plants in Germany, but they can also be more generic entities like 'refinery' and 'pipeline'. These entities can be extracted using an open-source toolkit such as NLTK (Loper & Bird, 2002), which can also be used to identify the types of entities mentioned, differentiating persons, locations, dates, amounts, etc. We could also enhance NLTK results by using transfer learning techniques to make it possible to identify finance-specific people and places (Lee et al., 2017) as well as using external resources such as DBPedia to extend the scope of the entity extraction (Auer et al., 2007).

Furthermore, in order to glean additional information from the climate disclosures, we propose using the Open IE algorithm (Banko et al., 2007), which can extract sets of relational tuples such as *the issue of climate change, could, reduce demand for our products*[2] in an unsupervised manner. However, since the Open IE approach is greedy and consequently produces a large number of output tuples, we aim to cross-reference its output with the named entities identified previously as well as the climate vocabulary used during the section ranking process. This will reduce the number of tuples extracted and allow us to identify the discourse accompanying specific terms such as 'climate change' and 'greenhouse gas', and to use sentiment analysis approaches to classify the discourse as positive or negative. Finally, since the type of vocabulary as well as the meanings of many words used in financial discourse differs from the vocabulary used in traditional sentiment analysis corpora, we propose to use financial word lists (Loughran & McDonald, 2011) to bootstrap existing sentiment analysis approaches (Wilson et al., 2005).

3. Projected Results

Using the methodology defined above, we aim to give investors, analysts and policy makers the tools they need to:

1. Find which companies are addressing the threat of climate change in their annual disclosures;
2. Narrow down the sections in financial disclosures that address climate-related risk;
3. Analyze what is being disclosed regarding various topics (e.g. emissions) or events (e.g. Paris Agreement);
4. Compare what companies from a given sector are disclosing or not regarding their climate liabilities;
5. Infer their long-term climate exposure of companies that have not publicly disclosed climate risk based on factors such as their industry, sector and geographical location.

In conclusion, we propose a combination of human intervention and NLP to carry out an in-depth analysis of the breadth and content of climate change in disclosures. This task is difficult for humans, but nonetheless a fully-automated NLP pipeline may not achieve the required granularity and precision. Adding human validation in the loop could make the results more transparent and the analysis more comprehensive. We foresee that ML tools can be closely combined with human skills to deal with other social issues beyond climate change.

[2] Taken from BP's 2017 annual report: https://www.bp.com/content/dam/bp-country/de_ch/PDF/bp-annual-report-and-form-20f-2017.pdf

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pp. 722–735. Springer, 2007.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. Open information extraction from the web. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, volume 7, pp. 2670–2676, 2007.

Doran, K. L. and Quinn, E. L. Climate change risk disclosure: a sector by sector analysis of SEC 10-k filings from 1995-2008. *NCJ Int'l L. & Com. Reg.*, 34:721, 2008.

Gamble, G. O., Hsu, K., Kite, D., and Radtke, R. R. Environmental disclosures in annual reports and 10Ks: An examination. *Accounting Horizons*, 9(3):34, 1995.

Hirji, Z. Most U.S. companies ignoring SEC rule to disclose climate risks. *InsideClimate News*, 2013.

Lee, J. Y., Dernoncourt, F., and Szolovits, P. Transfer learning for named-entity recognition with neural networks. *arXiv preprint arXiv:1705.06273*, 2017.

Loper, E. and Bird, S. NLTK: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

Loughran, T. and McDonald, B. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

Pennington, J., Socher, R., and Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

U.S. Securities and Exchange Commission. Commission guidance regarding disclosure related to climate change. 33-9106; 34-61469; FR-82, 2010.

Wilson, T., Wiebe, J., and Hoffmann, P. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.

Winn, M., Kirchgeorg, M., Griffiths, A., Linnenluecke, M. K., and Günther, E. Impacts from climate change on organizations: a conceptual foundation. *Business strategy and the environment*, 20(3):157–173, 2011.