4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15074

Collaborate for what: a structural topic model analysis on CDP data

Camilla Salvatore[1], Alice Madonna[2], Annamaria Bianchi[3], Albachiara Boffelli[2], Matteo Kalchschmidt[2]

[1]Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy, [2]Department of Management, Information and Production Engineering, University of Bergamo, Bergamo, Italy, [3]Department of Management, Economics and Quantitative Methods, University of Bergamo, Bergamo, Italy

Abstract

This paper aims to understand why firms engage with their suppliers to collaborate for sustainability. For this purpose, we use the Carbon Disclosure Project (CDP) Supply Chain dataset and apply the Structural Topic Model to: 1) identify the topics discussed in an open-ended question related to climate-related supplier engagement and, 2) estimate the differences in the discussion of such topics between CDP members and non-members, respectively focal firms and first-tier suppliers. The analysis highlights that the two prominent reasons why firms engage with their suppliers relate to several aspects of the supply chain management, and the services and good transportation efficiency. It is further noted that first-tier suppliers do not possess established capabilities and, therefore, are still improving their processes. On the contrary, focal firms have more structured capabilities so to manage supplier engagement for information collection. This study demonstrates how big data and machine learning methods can be applied to analyse unstructured textual data from traditional surveys.

Keywords: *sustainable supply chain management; carbon disclosure project; supplier collaboration; structural topic model; text mining*

1. Introduction

In recent years, environmental disclosure programs, in which firms communicate how they manage their impact on climate change, are gaining more and more traction. While at the beginning of the 2010s, these programs were deemed to provide a competitive advantage, today, they are almost mandatory in a supplier selection procedure (Serafeim, 2020). Previous studies have utilised data from these programs to understand their impact on the firm's performances (Madonna, Boffelli and Kalchschmidt, 2021), but they have failed to understand the reasoning behind different behaviours. Particularly, distinguishing between different tiers along the SC is crucial, as their approach to sustainability could have happened in different time frames and for different reasons (Schmidt et al., 2017). Thus, this study aims to fill this gap by trying to answer the following research question:

"Why do firms collaborate for sustainability along their supply chain?"

Taking on this goal, the Carbon Disclosure Project (CDP) Supply Chain (SC) dataset has been considered suitable due to the depth of information provided and the availability for the respondent to describe the engagement strategies through open-ended questions (CDP, 2018). The availability of open-ended survey questions allows us to deeply investigate the behaviour of businesses with respect to standard closed-ended questions. However, texts are unstructured data and Machine Learning (ML) approaches are fundamental to extracting information from such data. To this purpose, we apply the Structural Topic Model (STM) technique, which allows us to discover the latent topics discussed and to estimate the effect of relevant metadata (being a Member of CDP) on the discussion proportion of topics. The main reasons for joining CDP concern enhancing the firm's image and reputation and receiving insights into one's suppliers. The data are gathered thanks to CDP members who request their suppliers to fill in a questionnaire to report information about climate change management, after filling in the questionnaire themselves. We expect the comments to highlight a divergence in the reasoning behind the engagement from the firms that, leveraging the data collection procedure, we can allocate into different tiers in the supply chain. In particular, we assign to CDP Members the role of *focal firms*[1] and the Non-Members the role of *first-tier suppliers*.

The novelty of the work is to be found firstly in the methodology, which is approached in the field of Sustainable Supply Chain Management (SSCM) for the first time. Indeed, in the field of sustainability, STM has been applied only to study open-ended questions about climate change (Tvinnereim, & Fløttum, 2015) and CSR disclosure in tweets (Salvatore, Biffignandi & Bianchi 2020). The second novelty introduced by this work relies on the analysis's

[1] Focal firms are those firms considered the leaders and the power fulcrum of their supply chain. The distinction between focal firms and first-tier suppliers has been done by leveraging the data collection procedure.

perspective. The “business-as-usual” of SSCM research is to observe how business decisions impact firms’ performance. This study has taken on the challenge to invert the viewpoint, considering that sustainability actions are required, mandatory at times (Serafeim, 2020), even though they do not necessarily influence firms’ performances (Pagell & Shevchenko, 2014).

The remainder of this article is organised as follows. Section 2 introduces the model. In Section 3, the data and the model selection strategy are presented. The results are discussed in Section 4. The main conclusions are drawn in Section 5.

2. The Structural Topic Model (STM)

Topic modelling (TM) is an unsupervised learning technique that allows studying the underlying properties of a text to discover the topics discussed and get signals from the data (Vayansky & Kumar, 2020). Among the different algorithms to implement TM, we select the STM, which was originally designed to analyse open-ended survey questions, and which is becoming increasingly popular due to the possibility of estimating models including document-level metadata, thus characterising the relationship between topics and metadata (Roberts, Stewart, & Airoldi, 2016).

In the following, we briefly introduce the STM algorithm. Please refer to Roberts, Stewart, & Airoldi (2016) for more details. STM is based on the bag of words assumption, which means that each document is represented as a vector of words without considering the order in which they appear. A topic is defined as a mixture of words, and a document as a mixture of topics. In STM, document-metadata influences two components of the model, the topical prevalence, which is defined as the proportion of the document associated with a topic, and the topical content, which refers to the usage rate of a word in a topic. Thus, topical prevalence covariates affect the discussion proportion of the topic (θ), while topical content covariates affect the rate of word usage within a topic (β). Here we focus only on topical prevalence covariates. The model can be represented in plate notation as in Figure 1.

The first step in estimating the model is to specify the algorithm initialisation strategy and the number of topics. Usually, the output is very sensitive to the specified initialisation. In this respect, the suggestion is to use spectral initialisation, a deterministic algorithm based on the method of moments, due to its stability (Roberts, Stewart, & Tingley, 2019). With respect to the specification of the number of topics, it is worth noticing that there is no true number of topics, and the suggestion is to test different numbers of topics by comparing some metrics and manually evaluating the results. Roberts, Stewart, & Tingley (2019) argue that four metrics should be compared: held-out likelihood, residual dispersion, semantic coherence and exclusivity. The held-out likelihood is a measure of the predictive power. The higher the held-out likelihood, the higher the model’s predictive power. Residual dispersion is equal to one when the model is well specified. This is a very strict requirement, and for practical purposes, the analyst should prefer models with low residuals and evaluate residuals in combination with the other metrics. Semantic coherence measures the co-appearance rate of the most probable words in that topic, so the higher this metric is, the better a topic is defined. However, semantic coherence decreases as the number of topics increases, i.e., if the number of topics is small, it is likely that they will be composed of the same words. Thus, practitioners should also look at exclusivity, which measures whether the top words for that topic do not appear as top words in other topics (exclusivity of words to a topic).

After the initialization and the number of topics is specified, model estimation and inference are based on an appropriate variational E-M algorithm, which returns as output the discussion proportion of the topics for each document, the rate of word usage within each topic, and the effect of covariates on the topical prevalence and topical content.

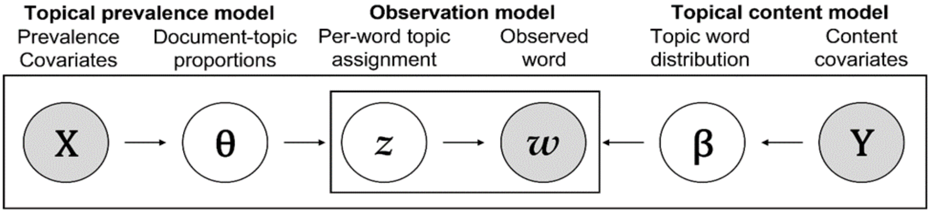

Figure. 1. Structural Topic Model. Source: Amended from Roberts et al. (2016).

For our analyses, we use R and, in particular, the `STM` package (Roberts et al., 2019) to estimate the model and the `quanteda` package (Benoit et al., 2018) to clean and prepare the data.

3. Data and model selection

The data we analyse are part of the SC dataset of the 2018 CDP questionnaire (CDP, 2018). The Carbon Disclosure Project is a non-profit organisation which encourages firms to disclose information about their climate-change-related risks and opportunities through a yearly survey. Figure 2 shows the sample refinement process, which lead to a final dataset of 314 firms. Each respondent could comment on different types of the deployed engagement strategies, namely Compliance & onboarding, Information collection, Engagement & incentivisation, Innovation & collaboration, and others. Thus, 461 short comments on the different rationale for why the 314 respondents engage their suppliers.

Before analysing the comments, it is necessary to clean the data. This involves different operations aiming at keeping only relevant words, reducing the complexity of the model and speeding up the estimation process. In this respect, the steps we implemented are: elimination of punctuation, stop words and numbers, conversion to lowercase, and stemming. An additional step in data cleaning is the removal of infrequent terms. This allows reducing the noise in the data, making topic detection easier. The rule of thumb is to remove the terms that appear in less than 0.5-1% of the documents (Denny & Spirling, 2018). We fixed the threshold to 1%, and the final set of unique stems is composed of 728 units. These data are ready to be analysed.

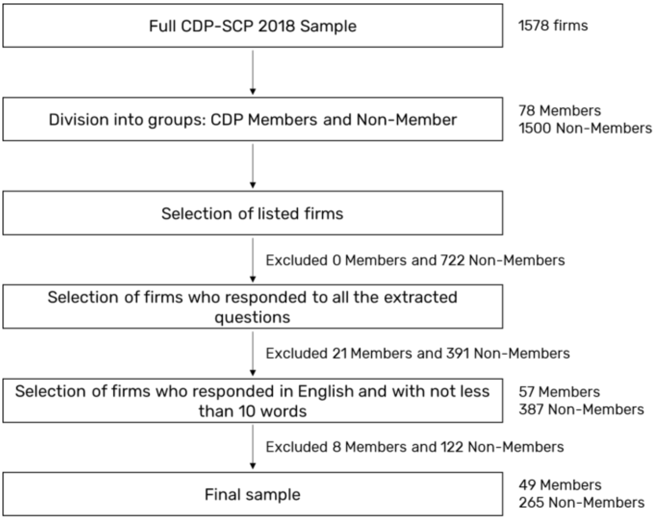
Full CDP-SCP 2018 Sample
1578 firms
Division into groups: CDP Members and Non-Member
78 Members
1500 Non-Members
Selection of listed firms
Excluded 0 Members and 722 Non-Members
Selection of firms who responded to all the extracted questions
Excluded 21 Members and 391 Non-Members
Selection of firms who responded in English and with not less than 10 words
57 Members
387 Non-Members
Excluded 8 Members and 122 Non-Members
Final sample
49 Members
265 Non-Members

Figure 2. Sample refinement procedure.

We consider only one topic prevalence covariate: being a member of the CDP. We study the effect of this variable on the proportion of discussion of topics.

The optimal number of topics is identified by looking at the metrics described in Section 2 and represented in Figure 3. Although it is not possible to identify the *true* number of topics, this procedure helps identify a set of plausible values. The appropriate number of topics seems to be around 20 and 30, where residuals are relatively low. After a manual evaluation of the quality of topics, we selected the model with 20 topics.

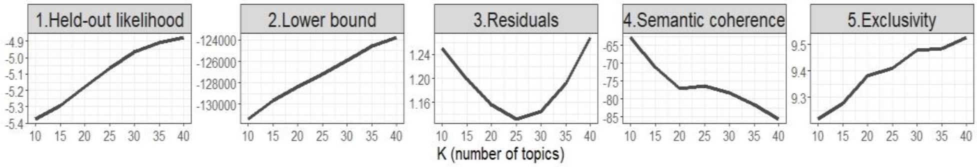

Figure 3. Evaluation metrics for choosing the number of topics

4. Results and discussion

Topics are identified by looking at the most probable stems for each topic, and labelling them consequently. Figure 4 shows the proportion of identified topics classified by macro-dimension (left panel) and how topics are correlated (right panel).

The most prevalent dimension (30% of the topics) relates to Supply Chain Management (SCM) issues, particularly suppliers' management (Topic 10, 13, 17), control (T5, 6) and accountability (T4). The second one (20%) is about Services and Materials Transportation (SMT), which addresses transport optimisation (T12), outsourcing of services (T9, 11), and transportation of sold goods (T3). The remaining dimensions relate to the measurement of GHG emissions and more globally carbon footprints (MS – T7, 8, 19), compliance with different standards (COMP – T1, 16, 18), the use of data to make informed decisions (DDE - T14, 15), and finally, activities to promote sustainability (PS – T2, 20). Figure 4 (right panel) clearly shows that these macro-topics are not independent of each other.

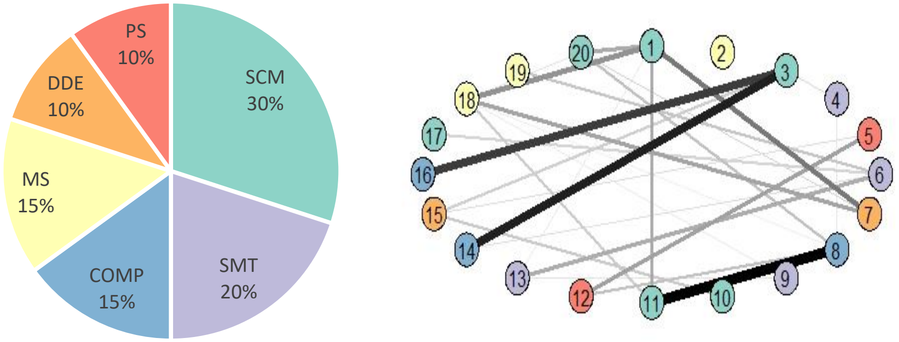

Figure 4. Left panel: Proportion of topics by macro-dimension (SCM: Supply Chain Management - green; COMP: Compliance - blue; SMT: Services and Materials Transportation - purple; MS: Measuring Sustainability - yellow; DDE: Data-Driven Evaluations – orange; PS: Promoting Sustainability - red. Right panel: Correlation plot.

Looking at the effects of being a CDP member on the discussion proportion of the topics, we estimate the changes in topic proportions shifting from firms that are CDP members and firms that are not. It turns out that differences are significant for 7 topics out of 20, as represented in Figure 5. In particular, topics 9, 11 and 12 that are prevalent for Non-Members refer to outsourcing services and optimising the firm's processes, all topics that concern an active process. Instead, Members are characterised by topics that refer to the management of other SC actors and the measurement of different parameters (topics 15, 18, 2 and 4).

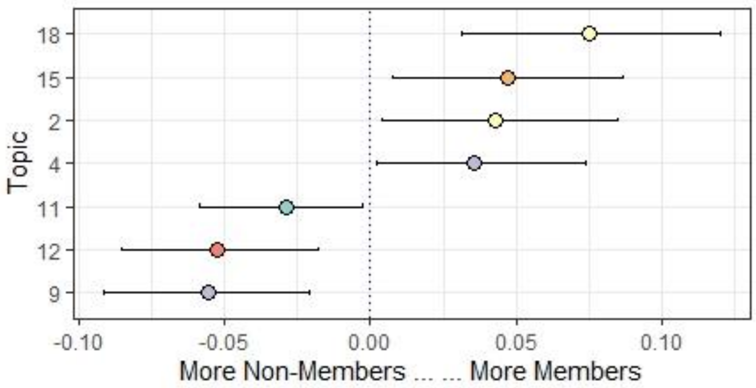

Figure 5. Estimated topic proportion difference between CDP members and non-members with 95% confidence interval.

5. Conclusions

In this work, we propose to apply the STM model for analysing one open-ended survey question about the rationale of engaging with suppliers to pursue sustainability-related goals along the supply chain. From a methodological point of view, although we do not consider a big data source, our research shows how machine learning approaches can be applied to unstructured textual data from traditional surveys to study socio-economic matters.

From a substantive point of view, implications of this work concern the enlightenment of the goals divergence for CDP members, namely focal firms, and non-members, namely first-tier suppliers, when it comes to collaborating along the supply chain for sustainability. This result supports the initial hypothesis to allocate the firms by membership into different tiers of the supply chain (focal firms and first-tier suppliers), as it supports the theoretical characteristics that belong to each category. For instance, focal firms have been classified as first movers toward the transition to sustainability, and therefore they have established resources and capabilities to confront the relevant stakeholders (Schmidt et al., 2017). On the contrary, first-tier suppliers are late entrants into the sustainability movement and, therefore, are still adapting their operations.

Future developments of the work foresee the development of an econometric model wherein the topic model will try to estimate how the discussion is reflected in the firm’s value and performance. The ultimate goal will be to provide managerial implications on how these environmental disclosure programs results are perceived and acknowledged over time by external parties.

References

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., & Müller, S. M. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software,* 3(30), 774. doi: doi: 10.21105/joss.00774

CDP. (2018). *Closing the Gap: Scaling up sustainable supply chains*. Retrieved from https://www.cdp.net/en/research/global-reports/global-supply-chain-report-2018/

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2), 168-189.

Elijido-Ten, E. O. (2017). Does recognition of climate change related risks and opportunities determine sustainability performance? *Journal of Cleaner Production, 141*, 956-966. doi:10.1016/j.jclepro.2016.09.136

Madonna, A., Boffelli, A., & Kalchschmidt, M. (2021). The role of sustainable supply chain management on improving environmental performance: a longitudinal analysis of CDP data. In *Proceedings of 52nd Annual Conference: Decision Sciences Institute,* ISBN: 978-0-578-62648-2, ISSN: 2471-884X,.

Pagell, M., & Shevchenko, A. (2014). Why Research in Sustainable Supply Chain Management Should Have no Future. *Journal of Supply Chain Management, 50*(1), 44-55. doi:10.1111/jscm.12037

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, 111, 988-1003.

Roberts, M., Stewart, B., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software,* 91(1), 1-40.

Salvatore, C., Bianchi, A., & Biffignandi, S. (2020). Communicating Corporate Social Responsibility through Twitter: a topic model analysis on selected companies. In *CARMA 2020: 3rd International Conference on Advanced Research Methods and Analytics*, pp 269-277

Schmidt, C. G., Foerstl, K., & Schaltenbrand, B. (2017). The Supply Chain Position Paradox: Green Practices and Firm Performance. *Journal of Supply Chain Management,* 53(1), 3-25. doi:10.1111/jscm.12113

Serafeim, G. (2020). Social-impact efforts that create real value. *Harvard Business Review, 98*(5), 37-48.

Tvinnereim, E., & Fløttum, K. (2015). Explaining topic prevalence in answers to open-ended survey questions about climate change. Nature Climate Change, 5(8), 744-747

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems, 94*, 101582.