# CARMA 2020

July 8-9, 2020        Valencia, Spain

## 3rd International Conference on Advanced Research Methods and Analytics

*Congress UPV*

3rd International Conference on Advanced Research Methods and Analytics (CARMA 2020)

## Scientific Editors

Josep Domenech
María Rosalía Vicente

## Publisher

# Communicating Corporate Social Responsibility through Twitter: a topic model analysis on selected companies

**Camilla Salvatore[1], Annamaria Bianchi[2], Silvia Biffignandi[2]**

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

## Abstract

*Social media are fundamental in creating new opportunities for firms and they represent a relevant tool for the communication and the engagement with customers. The purpose of this paper is to analyse the communication of Corporate Social Responsibility (CSR) activities on Twitter. We consider the listed companies included in the Dow Jones Industrial Average Index and we implement a topic model analysis on their timelines. In order to identify the topic discussed, their correlation, and their evolution over time and sectors, we apply the Structural Topic Model algorithm, which allows estimating the model including document-level metadata. This model proves to be a powerful tool for topic detection and for estimating the effects of document-level metadata. Indeed, we find that the topics are overall well identified, and the model allows catching signals from the data. Finally, we discuss issues related to the validity of the analysis, including data quality problems.*

*Keywords: Topic modelling; Structural Topic Model; Social media communication.*

## 1. Introduction

Social media are fundamental in creating new opportunities for firms and they represent a relevant tool for engaging with customers and stakeholder. Also the communication of corporate social responsibility (CSR) activities, which plays a fundamental role in enhancing firms' reputation, can enjoy the new opportunities deriving from their use (Cho, Furey, & Mohr, 2017). There is not a unique and shared definition of CSR. Table 1 provides an overview of the different classifications and shows that it is a multidimensional concept.

**Table 1. CSR dimensions.**

| Reference | Dimensions |
|---|---|
| Carroll (1991) | Economic, Legal, Ethical and Philantropic |
| Dahlsrud (2008) | Environmental, Social, Economic, Stakeholder and Voluntariness |
| Kim et al. (2014) | Environmental, Philantropy, Education, Community/Employee involvement, public health, sponsorship of cultural/sports activities |

Source: Amended from Cho et al. (2017).

Although computer-assisted analysis of CSR reports is common, the literature about the analysis of social media messages about CSR is scarce (Chae & Park, 2018). The purpose of this paper is to analyse the communication strategy of CSR activities through Twitter by a selected group of firms in order to answer the following questions.

**Question 1.** *Which CSR topics are discussed on Twitter?*

**Question 2.** *Which CSR topics are sector-specific?*

**Question 3.** *Which topics are likely to be discussed together?*

**Question 4.** *What is the topic evolution over time?*

The novelty of this paper lies in the following aspects. First, we focus on the messages posted by a selected group of companies rather than retrieving tweets that match a specific search query (group of relevant keywords). Second, for answering our questions, we apply the Structural Topic Model (STM) algorithm, which allows estimating the model including document-level metadata.

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)

The STM is a probabilistic mixed membership model which allows to estimate a model including document-level metadata and, thus, to study the relationship between topics and metadata. In this section, we briefly describe the model; for further technical details, please refer to Roberts, Stewart, & Airoldi (2016), which originally proposed the model. This model is based on the *bag of words* representation, which means that each document is represented as a vector of words without giving importance to the order in which they appear. Let us consider a set of $D$ documents indexed by $d \in \{1 \dots D\}$. Each document is composed by a mixture of words $w_{d,n}$, where $n \in \{1 \dots N_d\}$ indicates the position within the document. The collection of unique words is represented by a vocabulary. Each term in the vocabulary is indexed by $v \in \{1 \dots V\}$, it is assigned to a topic ($z$) and it is associated with the probability of belonging to each topic $k \in \{1 \dots K\}$. Thus, a topic is a mixture over words and the document is a mixture over topics. Document-metadata influence two components of the model, the *topical prevalence* that is defined as the proportion of the document that is associated to a topic, and *topical content* that refers to the usage rate of word in a topic. Thus, topical prevalence covariates affect the discussion proportion of the topic ($\theta$), while topical content covariates affect the rate of word usage within a topic ($\beta$). The matrix of the $P$ topic prevalence covariates and $A$ topical content covariates are denoted by $X_{D \times P}$ and $Y_{D \times A}$ respectively. Model estimation and inference are based on a collapsed variational expectation-maximization algorithm. The model converges when the relative change in the approximate variational lower bound is below a defined tolerance level. Figure 1 summarizes the STM and highlights its three components: the *topic prevalence model* (left-hand side), the *topical content model* (right-hand side), and the *observation model* (central part).
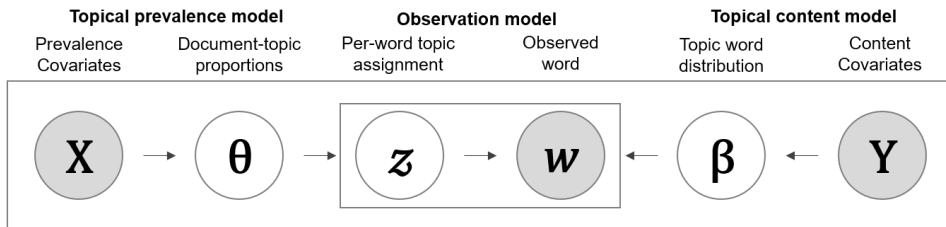


*Figure 1. Structural Topic Model. Source: Amended from Roberts et al. (2016).*

When estimating the model, the analyst must specify the algorithm initialization strategy and the number of topics. A shortcoming of topic models is that the output is very sensitive to the initialization. The *spectral initialization*, a deterministic algorithm based on the method of moments, is suggested due to its stability (Roberts, Stewart, & Tingley, 2019). Then, for choosing the optimal number of topics, it is necessary to compare some metrics. Roberts et al. (2019) argue that four metrics should be compared: residuals dispersion, held-out likelihood, semantic coherence and exclusivity. The held-out likelihood is a measure of

predictive power, which is useful for models comparison. The authors apply the document completion approach to estimate the held-out likelihood. The higher the held-out likelihood, the higher the model's predictive power. Taddy (2011) suggests that the dispersion of the residuals is one when the model is well specified. Residuals' dispersion is checked by means of a chi-squared test ($H_0: \sigma^2 = 1$ vs $H_1: \sigma^2 > 1$). A large number of topics should be preferred when rejecting the null. However, this requirement is very strict and for practical purposes, it is suggested to look at residual dispersion together with the other metrics. Mimno et al. (2011) present the concept of semantic coherence that is calculated for each topic $k$ and it provides a measure of the co-appearance rate of the most probable words in that topic. If the most probable words in the topics tend to co-occur, then the topic is semantically coherent. Let $V^{(k)} = \left(v_1^{(k)}, \dots, v_M^{(k)}\right)$ be the list of the $M$ most probable words in topic $k$. Then, define $D(v)$ as the *document frequency* for word $v$, and $D(v_m, v_l)$ as the *co-document frequency* for words $v_m$ and $v_l$, i.e., the number of documents in which the selected terms occur together. Then, for each topic $k$, the semantic coherence is defined as follows:

$$C\left(k; V^{\{(k)\}}\right) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D\left(v_m^{(k)}, v_l^{(k)}\right) + 1}{D(v_l^{(k)})}.$$

It is easy to check that semantic coherence will decrease as the number of topic increases, i.e., if the number of topics is small, it is likely that they will be composed by the same words. As countermeasure, Roberts et al. (2016) suggest to consider a measure of exclusivity, called FREX. Airoldi and Bischof (2016) develop this metric in a way that words frequency is balanced by exclusivity. Define $B(v^{(k)})$ as the *occurrence rate* of a word $v$ in topic $k$. Then, for a set of comparison topics $S$, the exclusivity is defined as follows $E(k; v) = B(v^{(k)})/\sum_{h \in S} B(v^{(h)})$. The FREX is defined for each topic $k$ and term $v$ as the weighted harmonic mean of term's frequency and exclusivity:

$$FREX_{k,v} = \left(\frac{w}{ECDF(B(v^{(k)})/\sum_{h \in S} B(v^{(h)}))} + \frac{1 - w}{ECDF(B(v^{(k)}))}\right)^{-1}$$

where $w$ is the weight in favour of exclusivity and ECDF stands for empirical cumulative distribution function.

## 3. Data and model selection

We selected the firms included in the Dow Jones Industrial Average index, i.e., a stock market index that measures the performance of the 30 largest US listed companies. We retrieved the full list of firms, joint with the activity sector from Bloomberg. Then, the original tweets (including retweets without a comment) posted on the firm's timeline have been collected. As reference period, we selected the second semester of 2019 (July-

December). Two firms (Apple and Walgreens) turned out not to have a Twitter account, while Walmart has been excluded due to rate limiting when retrieving data. The final sample includes 27 firms. Most of them operates in the Financial (18.5%), Technology (14.8%) and Health Care (14.8%) sectors. Then, there are Industrials and Consumer Discretionary (11.1% each) sectors, and Communication, Consumer Staples, Energy and Material ones (7.4% each). The number of messages retrieved is 8,602.

The stm R package developed by Roberts et al. (2019) has been used for implementing the analyses. The first step concerns the cleaning of the data. It involves different operations: elimination of punctuation, stop words, numbers, conversion to lower case, and stemming. The data are finally organized into documents, vocabulary terms and tokens (repeated words) as follows: 8,602 documents, 23,983 unique words and 136,201 tokens. After the cleaning process, only relevant terms remain. However, an additional step in data cleaning is the removal of infrequent terms (those that appear in a number of documents less or equal to a threshold). The threshold is defined as the number of documents in which the word appears. This operation is highly recommended because it allows reducing noise in the data, making the task of topic detection easier.

The choice of the appropriate threshold is made by looking at the number of the remaining documents, words and tokens (Table 2). Then, the analyst can assess the remaining terms in order to choose the appropriate threshold. For low values of the threshold, the reduction in the noise is small, thus we focused on higher values of thresholds, more specifically on 20, 30 and 50. After analysing the words that compose the vocabulary for each case, the most appropriate threshold seems to be 30.

**Table 2. Comparison of thresholds.**

| Threshold | No. Documents | No. Words | No. Tokens |
|:---:|:---:|:---:|:---:|
| 20 | 8591 | 1172 | 89663 |
| 30 | 8584 | 824 | 81024 |
| 50 | 8559 | 503 | 68486 |

Source: Authors' own elaboration.

The next steps are model specification and identification of the optimal number of topics. In our analysis, we only include topic prevalence covariates. We allow sectors and day to affect the discussion proportion of a topic. We estimate the day variable through a spline in order to account for non-linear effects. The optimal number of topics is chosen by looking at the metrics described in Section 2 (Figure 2, left-hand side). The appropriate number of topics seems to be around 40 and 50. It should be clear that there is no fixed way to choose among them, and this procedure does not yield the *true* number of topics. The differences in terms

of held-out likelihood and residuals dispersion are small. The trade-off between semantic coherence and exclusivity is evident. In order to choose among them, Figure 2 (right-hand side) compares the two metrics. The 32.5% of the 40 topics falls in the first quadrant, the 55% in the second one and the 12.5% in the fourth one. For the model with 50 topics, the percentages are 30%, 60%, 10% respectively. Thus, the model with 40 topics seems to be most appropriate.

## 4. Results and discussion

Topic discovery is performed by looking at the most-probable words for each topic, and labeling them consequently. We identify 21 topics related to CSR activities (Figure 3 left-hand side).
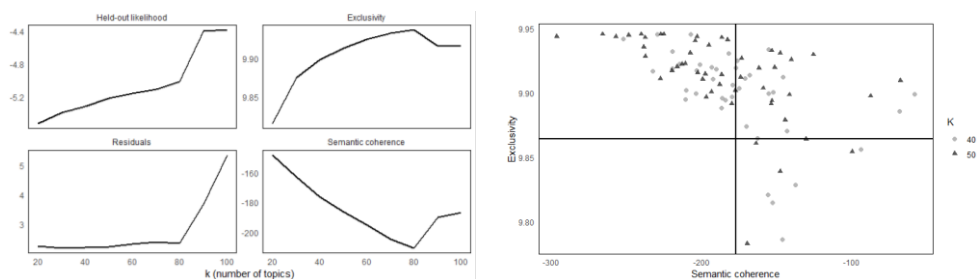


*Figure 2. Evaluation metrics for choosing the number of topics (left) and comparison between exclusivity and semantic coherence for models with 40 and 50 topics (right). Source: Authors' own elaboration.*

More precisely, 24.8% of them concerns the social dimension (community, employee engagement and sponsorship of events), 14.6% relates economic matters, 3.13% is on public health commitment, 2.13% concerns the environmental question, and finally, 1.95% of messages relates to educational programs.

The topics proportion of the identified CSR topics is plotted on the left-hand side of Figure 3. The topic correlation network is plotted on the right-hand side. It shows positively correlated topic, i.e., those topic that are likely to be discussed together within a tweet. Only correlations whose value is greater than 31% are plotted. Correlations within the same dimensions are evident. Moreover, two clusters have relevant features. Topics 39 (Education), 20 (Social), 25, 26 and 9 refers all to technological aspects. Topic 39 relates study programs involving technological instruments, Topic 20 relates the "digital transformation for helping communities" while the other topics are about the release of technological products or advertising about artificial intelligence, machine learning and digital services. The second cluster which includes "non CSR topics" (Topics 40, 19, 16, 33, 34, 1 and 17) relates advertising and promotions mainly linked to Christmas, Halloween and Summer holydays.
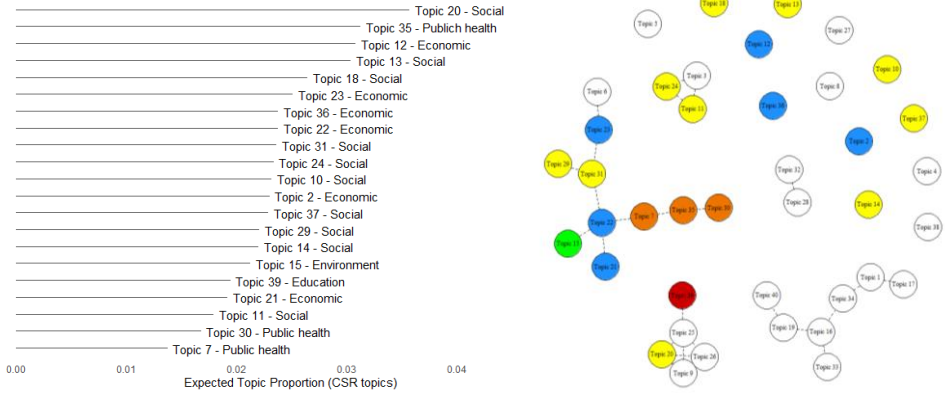
*Figure 3. Expected topic proportions of CSR topics (left) and topics correlation (right) with sector indication (Green for Environment, Red for Education, Yellow for Social, Blue for Economics, Orange for Public health).*

The novelty introduced by this model is the possibility of estimating the effects of topical prevalence covariates on the discussion proportion of a topic. We start from the sector variable (Figure 4). It is not surprising that firms in the energy and materials sectors tweet significantly more about environmental issues than the others (Topic 15). Interesting patterns can be observed for Topic 37 that concerns events sponsorship, mainly of the *#voteyourmainstreet* initiative. This event was sponsored by American Express, which belongs to the financial sector, i.e., the one that tweeted significantly the most. The second topical prevalence covariate is time. Figure 5 shows the expected topic proportion as a smooth function of the day with 95% confidence intervals. Topic 15 remained stable over time, with a small reduction during summer and winter holidays. Topic 37 shows a higher proportion during October and November, the months when the sponsored event mainly took place. Finally, Topic 10 that concerns supporting small businesses has a peak in the last days of November and the beginning of December. Indeed, in that period the *Small Business Saturday* initiative took place, that is a traditional event to support small businesses and for celebrating communities.
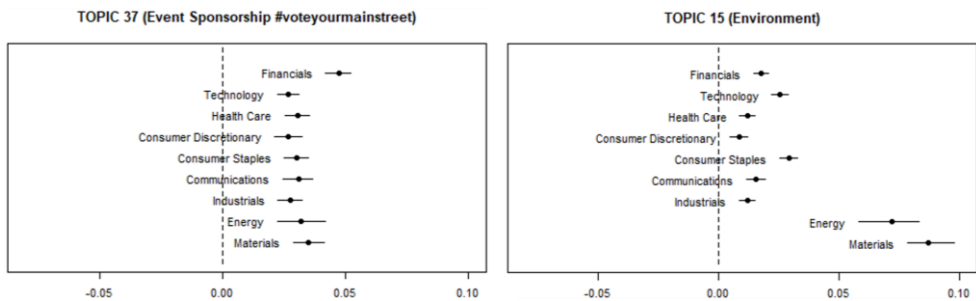


*Figure 4. Effect of the "sector" on the proportion of topic discussion. Source: Authors' own elaboration.*
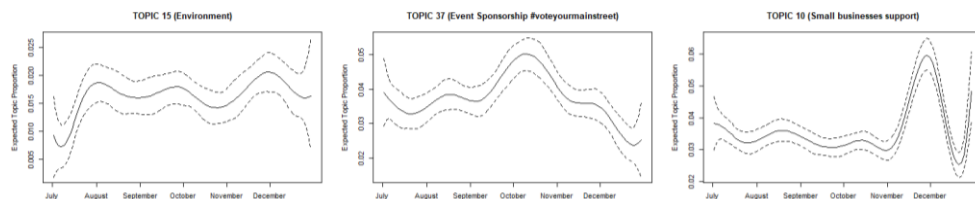
*Figure 5. Expected topic proportion over time. Source: Authors' own elaboration.*

## 5. Conclusions

In this work, we propose to apply the STM model for analyzing the communication of a selected group of firms about CSR activities on Twitter, allowing topical prevalence to evolve over time and varying across sectors. With reference to the initial questions, STM proves to be a powerful tool for topic detection and for estimating the effects of document-level metadata. Indeed, we get evidence that some topics are sector-specific and that the model allows to catch signals from the data, in correspondence of particular events. In addition, interesting correlations have been highlighted. When analyzing Twitter data, practitioners should be aware about data quality aspects and the errors they may encounter (Salvatore, Biffignandi, & Bianchi, 2020). Indeed, the main shortcoming is that the output of the analysis is very sensitive to the analyst' judgements at the various steps. Further developments may concern the analysis of data quality aspects, the inclusion of covariates' interactions, and of topical content metadata.

## References

Airoldi, E., & Bischof, J. (2016). Improving and Evaluating Topic Models and Other Models of Text. *Journal of the American Statistical Association, 111*, 1381-1403.

Carroll, A. (1991, 7 1). The pyramid of corporate social responsibility: Toward the moral management of organizational stakeholders. *Business Horizons, 34*(4), 39-48.

Chae, B., & Park, E. (2018, 6 28). Corporate Social Responsibility (CSR): A Survey of Topics and Trends Using Twitter Data and Topic Modeling. *Sustainability, 10*(7), 2231.

Cho, M., Furey, L., & Mohr, T. (2017). Communicating corporate social responsibility on social media: Strategies, stakeholders, and public engagement on corporate facebook. *Business and Professional Communication Quarterly, 80*(1), 52-69.

Dahlsrud, A. (2008, 1 1). How corporate social responsibility is defined: an analysis of 37 definitions. *Corporate Social Responsibility and Environmental Management, 15*(1), 1-13.

Kim, S., Kim, S., & Sung, K. (2014, 10 28). Fortune 100 companies' Facebook strategies: Corporate ability versus social responsibility. *Journal of Communication Management, 18*(4), 343-362.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & Mccallum, A. (2011). *Optimizing Semantic Coherence in Topic Models.* Association for Computational Linguistics.

Roberts, M., Stewart, B., & Airoldi, E. (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, 10 31). Stm: An R package for structural topic models. *Journal of Statistical Software, 91*(1), 1-40.

Salvatore, C., Biffignandi, S., & Bianchi, A. (2020). Social Media and Twitter Data Quality for New Social Indicators. *Social Indicators Research*. doi:10.1007/s11205-020-02296-w

Taddy, M. (2011, 9 21). On Estimation and Selection for Topic Models. *Journal of Machine Learning Research, 22*, 1184-1193.