Text mining and dimension reduction method application into exploring isomorphic pressures in corporate communication on textual tweet data about sustainability in the energy sector.

Atsuho Nakayama *Tokyo Metropolitan University* e-mail: *atsuho@tmu.ac.jp*
Emilia Smolak-Lozano *University Of Malaga* e-mail: *emilia.smolak@gmail.com*
Adriana Paliwoda Matiolańska *Cracow University of Economics* e-mail: *paliwoda@uek.krakow.pl*

The study analyses the isomorphism pressures within the context of sustainability by exploring the Twitter communication in the energy sector. Recently, there can be observed the increasing focus on interactive and communicative construction of an institution to understand how the organizations sustain the institutional pressures. The rhetorical commitments that create narrative dynamics in organizational communication are central to institutional diffusion and change. Social Media, Twitter, in particular, has been demonstrated as the new opportunity to explore the linguistic dimension in corporate communications. We propose the use of Social Media linguistic data (tweets with their hashtags and keywords) and the triangulated method (text mining, web mining, and linguistic and content analysis) to examine the tweets´ trends in each company. Based on the institutional theory of organizational communication, the paper examines the relation between the idea of sustainability and isomorphism that leads to the adoption of similar models and attitudes among the organizations. It applies the text mining and correspondence methods within the R software. The energy sector tweets in English (from 2016) were treated by the text mining processes of the statistical linguistic analysis in the R tool. Text mining, involving the linguistic, statistical, and the machine learning techniques reveals and visualizes the latent structures of the content in an unstructured or weakly structured text data in a given collection of documents. The method helps to represent the topic of a textual document containing a sample of tweets through the frequency study of the semantically significant terms used in these tweets. Document-term matrix has been calculated via text mining technique against the tweet data, then by aggregating it for each company and representing a word frequency in each company. Since the matrix is sparse and large, it has been necessary to perform a dimensionality reduction analysis to uncover the underlying semantic structure.

Dimensionality reduction methods such as Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, and Non-Negative Matrix factorization have been found to perform well for this task. Latent Semantic Analysis reduces the dimensionality of the document-term matrix by applying a singular value decomposition, and it then expresses the result in an intuitive and comprehensible form. In the Probabilistic Latent Semantic Analysis, a probabilistic framework is combined with the Latent Semantic Analysis. It has been shown that the Non-Negative Matrix Factorization and Probabilistic Latent Semantic Analysis alike optimize the same objective function, ensuring the equivalent use of both. The Non-Negative Matrix Factorization includes the positive coefficients in the linear combination. The computation is based on a simple iterative algorithm, which is particularly useful for applications involving a complicated linguistic tweets´ matrix. By the results of the analysis, we have clarified the tendency of words used by each company in their tweets, being able to determine the degree of homogeneity in the textual contents of the tweets. The results show the tendency among the energy companies to follow similar patterns in Twitter communication on sustainability. Therefore, we can observe the mechanisms that lead to isomorphism in organizational communication.