Analysing societal impacts of cultural policies and practices: A Business Intelligence Toolkit based on Natural Language Processing

Sanda Martinčić - Ipšić and **Božidar Kovačić** (University of Rijeka), **Francesco Molinari** (*University of València*)

Highlights

- The MESOC Toolkit is a georeferenced visualization tool and a semantic search engine helping to map and analyse the societal impacts of cultural policies and practices described in academic papers and grey (policy) literature texts.
- The AI was trained on a repository of free and open-source documents and registered users can upload new ones.
- Language is never a problem, thanks to the use of the Google API for translation from and to English, which is the language used by the Toolkit.
- The areas of societal impact are those identified by the EU Agenda for Culture as 3 "crossover themes": health and well-being, urban renovation and regeneration, social cohesion and public participation.
- The contribution of cultural activities is defined according to the 10 domains of the UNESCO Framework of Cultural Statistics.
- Two business cases have been identified so far: a researcher wanting to retrieve meaningful examples of societal impact of specific cultural domains; and a policy analyst willing to find similar documents to one or more already in his/her possession.
- The MESOC Toolkit is capable of automatic inference of impacts by means of projection into the semantic space, semantic similarity search and clustering.
- The URL is: https://toolkit.mesoc-project.eu/.

Background

The main objective of Business Intelligence (BI) is to facilitate decision making by integrating key information from various data sources - internal or external, structured or unstructured [Aramburu, Berlanga & Lanza, 2020]. BI has traditionally focused on analysing internal data, most of which stored in structured relational databases. Nowadays, BI must also deal with external and unstructured data such as texts from documents, web pages, social networks, short messages, user-generated ratings and comments, sensor data, etc. [Chung & Tseng, 2012; Sreesurya et al., 2020]. Therefore, in addition to traditional, multidimensional data processing techniques, modern BI requires the integration of Artificial Intelligence (AI) methods, especially for processing unstructured data.

Natural Language Processing (NLP) is a subfield of AI that aims to process and ultimately understand natural language in written or spoken form, hence unstructured data. To that end, NLP combines methods from computer science – such as machine and deep learning – but also statistics and linguistics or computational linguistics to solve various tasks: from keyword extraction [Beliga et al., 2015; 2016] to topic modeling [Bogović et al., 2021]; from text classification [Martinčić-Ipšić et al. 2019] to automatic text summarization [Aljević et al., 2021]; from sentiment analysis [Babić et al., 2021] to named entity recognition [Beliga et al., 2021], etc. Hence, the MESOC toolkit is a BI tool powered by NLP and AI methods [Bogović et al, 2022a, Bogović et al, 2022b].

Rationale

Both the academic and grey literature in all topic areas and specifically within the socio-economic field, usually bring with them insights taking the form of (e.g.) case study descriptions and interpretations. Arts and culture make no exception, and this is particularly true for a recently emerged research (and policy) strand focused on societal impacts of cultural events or activities [Matarasso, 1997; European Commission, 2018].

Notoriously, AI and NLP if adequately structured can help simplify the task of scientists and policy analysts. In MESOC, two business cases have been identified so far:

- A researcher on the societal impacts of culture would like to create and analyse a set of documents dealing with the influence of a specific cultural domain (e.g. arts and crafts) on a predefined impact area (e.g. urban renovation and regeneration);
- A policy analyst preparing a City's candidature as European Capital of Culture is searching for similar documents to the one(s) already in his/her possession, in order to complete an overview of previous applications and gain additional insights from available case descriptions in literature.

Solution

The MESOC Toolkit is based on a pipeline of NLP tasks using a repository of scientific and policy documents themed with arts and culture [Bogović et al, 2022a]. Processing of the documents begins with conversion from PDF (Portable Document Format) to text file type. Next, the language used is determined and translated into English using the Google API for machine translation.

Unsupervised keyword extraction is the central part of the NLP pipeline. The set of extracted keywords contains the most salient information from processed texts and serves as input for multi-label text classification and semantic search based on Jaccard similarity. A prerequisite for document analysis is the determination of the societal impact areas and cultural domains dealt with. This problem is formulated as a multi-label classification task in 30 classes (10 cultural domains from the UNESCO Framework of Cultural Statistics x 3 "crossover themes" from the EU Agenda for Culture). To mitigate the consequences of a too small number of database items, we adopted a non-standard approach to multiclass classification of texts: first, we classify the document according to the 3 societal impact areas, then we perform a classification into 10 domains, and finally we multiply the obtained class probability vectors to calculate the final class belonging probabilities.

However, the core functionality of the Toolkit is detection and analysis of societal impacts. At this final stage of the NLP pipeline, we apply an impact generation and semantic expansion method defined in [Bogović et al. 2022b]. The

method starts with the initial definition of 100 possible societal impacts defined by a domain expert, and expands their semantic neighbourhoods using word2vec embeddings, cosine similarity, and K-means clustering. The method creates a list of similar extracted n-grams from candidate texts and calculates a semantic similarity for each impact using a cosine similarity measure. K-means clustering is employed, aiming at avoiding overlaps between their corresponding semantic expansions. For each new document, the method extracts all bigrams and trigrams and matches them with n-grams in the semantic expansion of different impacts. For each match, the method detects the corresponding impact for the document and annotates it. Hence, the Toolkit is capable of automatic inference of impacts from the user documents.

Benefits from usage

The MESOC Toolkit is freely accessible from the following URL: https://toolkit.mesoc-project.eu/. It includes both a georeferenced visualization tool and a semantic search engine, allowing to:

- Visualise the geographical distribution of the case studies gathered in the repository.
- Attribute to each document the most relevant domains and impact areas in the form of a heat map.
- Suggest similar documents to the one under inspection that belong to the same cultural domain(s) or have impact on the same crossover theme(s).
- Access the full text(s) if physically stored in the underlying repository (Note: all of them are free and open source, or fully licensed).
- Derive possible impacts in the 3 crossover theme(s) of the EU Agenda for Culture.
- Registered users can perform the above analyses starting by uploading new documents.

References

Aljević, D.,. Todorovski, L., and Martinčić-Ipšić, S.: Extractive Text Summarization Based on Selectivity Ranking. In 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), pp. 1-6. IEEE (2021)

Aramburu, M.J., Berlanga, R., and Lanza, I.: Social media multidimensional analysis for intelligent health surveillance. International journal of environmental research and public health 17.7: 2289 (2020).

Babić, K., Guerra, F., Martinčić-Ipšić, S., and Meštrović, A.: A Comparison of Approaches for Measuring the Semantic Similarity of Short Texts Based on Word Embeddings. Journal of information and organiza-tional sciences, 44(2), 231-246 (2020)

Beliga, S., Meštrović, A., and Martinčić-Ipšić, S.: An overview of graph-based keyword extraction methods and approaches. Journal of information and organizational sciences 39.1. (2015)

Beliga, S., Meštrović, A., and Martinčić-Ipšić, S.: Selectivity-based keyword extraction method. International Journal on Semantic Web and Information Systems (IJSWIS) 12.3. 1—26 (2016)

Beliga, S., Martinčić-Ipšić, S., Matešić, S., Vuksanović, I.P., and Meštrović, A.: Infoveillance of the Croatian Online Media During the COVID-19 Pandemic: One-Year Longitudinal Study Using Natural Language Processing. JMIR public health and surveillance, 7(12), e31540 (2021)

Bogović, P.K., Meštrović, A., Beliga, S., and Martinčić-Ipšić, S.: Topic modelling of Croatian news during COVID-19 pandemic. In 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), pp. 1044-1051. IEEE (2021).

Bogović, P.K. Molinari, F. Kovačić, B. Martinčić-Ipšić, S. Generation and Semantic Expansion of Impacts in Arts and Culture // Advances in Information and Communication, Proceedings of the 2022 Future of Information and Communication Conference (FICC), Vol(1). San Francisco, USA: Springer International Publishing, 76-94, 2022b doi:10.1007/978-3-030-98012-2_8

Bogović,P.K., Aljević, D. Kovačić, K. Martinčić-Ipšić. S. The NLP Powered BI Toolkit: The Case of MESO. IEEE MIPRO 2022 - DE-DS (MIPRO'22), pp. 1348-1353, 2022a.

Chung, W., and Tseng, T. L. B.: Discovering business intelligence from online product reviews: A rule-induction framework. Expert systems with applications 39.15: 11870-11879 (2012).

European Commission (2018) A new European Agenda for Culture. https://ec.europa.eu/culture/sites/default/files/2020-08/ swd-2018-167-new-european-agenda-for-culture_en.pdf.

Martinčić-Ipšić S., Miličić T., Todorovski L.: The Influence of Feature Representation of Text on the Performance of Document Classification. Applied Sciences. 9(4): 743 (2019).

Matarasso, F.: Use or Ornament? The social impact of participation in the arts. Comedia (1997)

Sreesurya, I., Rathi, H., Jain, P., and Jain, T. K.: Hypex: A Tool for Extracting Business Intelligence from Senti-ment Analysis using Enhanced LSTM. Multimedia Tools and Applications, 79, (2020).

UNESCO Framework for Cultural Statistics (FCS) (2009). http://uis.unesco.org/sites/default/files/documents/measuring-cultural-participation-2009-unesco-framework-for-cultural-statistics-handbook-2-2012-en.pdf

Acknowledgments and disclaimer

This research has received funding from the European Union's (EU's) Horizon 2020 Research and Innovation Programme under Grant Agreement No. 870935. However, the opinions expressed here are solely of the authors and do not engage any of the EU granting institutions.

Technical overview

The MESOC Toolkit is implemented as a web application structured as a React application (the frontend), and as a REST API (the backend) written in Django Rest Framework that provides functionality to the frontend. In this way, the core backend functionality is decoupled from the frontend and can be used by other applications. Documents are parsed by running them through an NLP pipeline, where at each step a worker process performs a specific task (e.g. keyword extraction, impact generation, etc.). The worker processes are local to the backend, and there can be multiple instances running. Communication with the Django application is done through the Redis database, which serves as a message queue.

Frontend application is written in Javascript using the Node.js runtime environment, enabling to run javascript code outside of the browser. The application consists of two parts: The Nodejs + Expressjs based server and the React-based SPA (single page application). This design pattern enables us to have the development workflow and power of SPA while having the SEO (Search Engine Optimization) advantages for the homepage.

Semantic search is based on matching terms in query with keywords extracted from documents. The tool retrieves documents according to their semantic relatedness with the query terms according to Jaccard similarity measure. The proposed NLP method starts with an initial set of impacts and expands their semantic neighbourhood utilizing continuous space word representations - word2vec and cosine similarity measures.