The NLP Powered BI Toolkit: The Case of MESOC

Petar Kristijan Bogović[1,2], Dino Aljević[1,2], Božidar Kovačić [1], Sanda Martinčić-Ipšić[1,2]
[1]Faculty of Informatics and Digital Technologies, [2]Center for Artificial Intelligence and Cybersecurity,
University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia;
Email: {petar.kristijanb, dino.aljevic, bkovacic, smarti}@uniri.hr

***Abstract*—In this paper, we present a business intelligence (BI) toolkit based on natural language processing (NLP) methods for the arts and culture domain, collected in the Measuring the Social Dimension of Culture (MESOC) project. The main NLP methods in the underlying pipeline are keyword extraction, multi-label classification of texts, and detection of potential social impacts of cultural policies and practices, all trained on texts from open-access academic publications. The MESOC Toolkit is a georeferenced visualization tool for analyzing impact on social value creation in the areas of health and well-being, urban regeneration, and social cohesion, and enables semantic search for content in the MESOC domain. Therefore, the presented research can serve as a prototype for measuring societal value by identifying recurrent pathways of transformational processes in society that reach beyond the selected field of art and culture.**

Keywords - societal impact, natural language processing, business intelligence, keyword extraction, text classification, semantic search

I. Introduction

The main objective of business intelligence (BI) is to enable informed decision making by integrating strategic information from different data sources - internal or external, structured or unstructured [1], [2]. BI has traditionally focused on analyzing internal data for business analysis, most of which is stored in well-structured relational databases [3]. Today, in the wake of the data deluge caused by the big data phenomenon, BI must also consider external and unstructured data such as documents, web pages, social networks, short messages, user-generated ratings and comments, sensor data, etc. [4], [3], [2], [5]. Therefore, in addition to traditional multidimensional data processing, modern BI requires the integration of advanced artificial intelligence methods, especially for processing unstructured data.

Natural language processing (NLP) is a subfield of artificial intelligence that aims to process and ultimately understand natural language in written or spoken form [6], hence processes the unstructured data. NLP combines methods from computer science - machine and deep learning, statistics and linguistics or computational linguistics [6] to solve various tasks: keyword extraction [7], [8], topic modeling [9], text classification [10], automatic text summarization [11], sentiment analysis [12], named entity recognition [13], etc.

In this paper, we present the BI toolkit powered by the NLP pipeline for the arts and culture domain captured in the Measuring the Social Dimension of Culture (MESOC) project. The core of the project is MESOC Toolkit, based on NLP models trained on texts from open source academic papers published in peer reviewed journals or at international conferences that address the societal impacts of cultural policies and practices [14].

This paper is structured as follows: Section II provides an overview of the background and related work; Section III explains the NLP methods used in the NLP pipeline; Section IV presents the implementation details of the MESOC Toolkit web application with an overview of the documents used, while concluding remarks and directions for future research are presented in Section V.

II. Background and Related work

The Measuring the Social Dimension of Culture (MESOC) project is a Horizon 2020 Research and Innovation Action that aims to propose, test and validate an innovative and original approach to measuring the societal value and impact of culture, cultural policies and cultural practices [15]. Three main categories of societal impact are defined as follows: **impacts** on personal well-being, including improvements in physical and mental health; **impacts** on communities and social cohesion, associated with people's more intensive participation - both "passive" and "active" - in cultural activities; and **impacts** on urban regeneration and transformation, triggered by private and/or public investment in cultural assets (tangible or intangible) such as museums and exhibitions. The decision to choose these three impact categories stems from the New European Agenda for Culture [16], which explicitly focuses on the so-called "crossover processes" combining knowledge and skills specific to the cultural and creative sectors with those of other sectors in order to develop innovative and intelligent solutions to today's societal challenges, as originally outlined in the conclusions of a thematic EU Council (2015) [17], [18].

The focus of this research is placed around the 10x3 MESOC matrix, which cross-references three main social impact domains (Health and Well-being; Urban Renovation; Social Cohesion or People's Engagement and Participation) as three columns of the matrix to the cultural sectors. The rows of the matrix list ten cultural sectors that were included in the 2009 UNESCO Framework of Cultural Statistics [19] - which were partially, but not fully, included in the 2012

ESSnet-Culture Framework report by EUROSTAT [20]. The cultural sectors cover the main areas in which arts and culture operate, namely: 1. Heritage, 2. Archives, 3. Libraries, 4. Book and Press, 5. Visual Arts, 6. Performing Arts, 7. Audiovisual and Multimedia, 8. Architecture, 9. Advertising and 10. Art Crafts.

Therefore, the central goal of this research, is to provide policymakers and decision-makers with a tool to gain insight into the potential societal impacts of arts and culture initiatives, which are organized into three domains of the New European Agenda.

The MESOC Toolkit was developed using a selection of academic literature related to arts and culture that reflected the three categories of interest. For example, the link between arts and culture and health and well-being was explored starting with the seminal work of [21]. The interplay between urban and territorial regeneration and renovation and arts and culture gained a lot of attention, also starting from the groundbreaking European Capitals of Culture initiative [22]. Finally, increased social cohesion through the promotion of consumer/prosumer engagement and participation is also supported by the Council of Europe's Indicator Framework on Culture and Democracy [18].

III. Methods

The MESOC Toolkit is based on an underlying pipeline of NLP tasks for the arts and culture domain that are captured in the MESOC project documents. Processing of the documents begins with conversion from PDF (Portable Document Format) to text file type. Next, we determine the language of the text and translate it into English using the Google API for machine translation. Next, we remove cited references and clean up the noise introduced by the PDF to text format conversion. Unsupervised keyword extraction is the central part of the NLP pipeline, as elaborated in subsection III-A. The set of extracted keywords contains the most salient information from the processed texts and serves as input for the multi-label text classification described in III-B and the similarity search described in III-C. Finally, the core functionality for detection and analysis of derived impacts is elaborated in III-D.

A. Keyword Extraction

After preprocessing the text file, the Python library *pke* was used to extract keywords from the text. Pke is an open-source Python-based keyphrase extraction toolkit which provides an end-to-end keyphrase extraction [23] containing multiple different supervised and unsupervised keyphrase extraction models, such as tf-idf, YAKE [24], TextRank [25], and KEA [26]. In this work, we use an unsupervised keyphrase extraction model - YAKE [24]. Yet Another Keyword Extractor (YAKE) [24] uses statistical features of the single text document to identify the most important words - keywords [7], [8]. The algorithm consists of five steps: preprocessing the text and identifying term candidates, extracting features, computing the score for terms, generating n-grams and computing the score for term candidates, deduplicating data, and finally ranking the keywords.

In the first step, the algorithm applies typical preprocessing procedures such as cleaning the text, dividing it into sentences, annotating the text, tokenizing it, and identifying and removing stop words. First, the text is divided into sentences and then into chunks formed by tokens within each sentence, where each token is converted to lowercase and annotated depending on whether the token is a digit, unparsable content, acronym, uppercase letter, or other parsable content. In the feature extraction step, the algorithm iterates through the sentences and their corresponding chunks and splits the chunks into annotated tokens. For each token, a set of five features is computed. These features are: T_{case}, which reflects the capitalization of the term, giving higher importance to terms in uppercase form, $T_{position}$, which reflects the position of the term within the sentence, giving higher importance to terms that occur at the beginning of the sentence, TF_{norm}, which focuses on the frequency of the term within the document, but where the value of the term frequency is divided by the mean of the frequencies plus one times their standard deviation, with the goal of preventing bias toward higher frequencies in long documents. The next feature is T_{Rel}, which is computed depending on the terms surrounding the observed token. Here, terms that coexist with a higher number of unique tokens in the observed window are assigned lower values. The final feature is $T_{sentence}$, which indicates how often a candidate term occurs in different sentences. In the next step, for each term and its corresponding features, a unique score $S(t)$ is computed using the following formula:

$$\frac{T_{rel} \cdot T_{position}}{T_{case} + \frac{TF_{norm}}{T_{rel}} + \frac{T_{sentence}}{T_{rel}}}. \tag{1}$$

After computation, the algorithm creates a list of keyword candidates by generating a contiguous sequence of terms up to n-grams over a sliding window of n. For each candidate term kw, the following score is assigned:

$$S(kw) = \frac{\Pi_{t \in kw} S(t)}{KF(kw) \cdot (1 + \sum_{t \in kw} S(t))} \tag{2}$$

where kw represents a candidate keyword of one or more terms and $S(kw)$ represents the final score. $KF(kw)$ represents the frequency of the candidate keyword. The lower the $S(kw)$ score, the more relevant the candidate keyword candidate is.

After determining the ranking of candidate keywords, the final step of data deduplication and ranking is performed. The candidate keywords are added to a new list depending on their relevance. After each candidate keyword is added, it is semantically compared with each more suitable candidate keyword that was previously added to the list. If the semantic similarity is above a certain threshold, which means that this keyword has already been recognised, the keyword candidate is discarded. After manually reviewing the results, we decided to extract the top 30 4-gram keywords with a window of 18 and a deduplication threshold of 0.95.

B. Multi-label Classification

A prerequisite for document analysis is the determination of the societal impact domains and cultural sectors of the document, i.e., the columns and rows of the MESOC matrix. This problem is formulated as a multi-label classification task in 30 classes (10 cultural sectors and 3 domains of the MESOC matrix). Multi-label classification is inherently a challenging problem, but the difficulty is further compounded by the relatively small number of documents in the dataset [6], [27], [10]. To minimize this problem, we have chosen a non-standard approach to multiclass classification of texts: First, we classify the document into 3 cultural sectors, then we perform a classification into 10 domains, and finally we multiply the obtained class probability vectors to calculate the final class probabilities. In this way, we reduced the scale of multiclass classification into 30 classes to a simpler classification with 3 and 10 classes respectively.

Formally, let $P_r \in \mathbb{R}^{10}$ be the vector of probabilities for each domain and $P_c \in \mathbb{R}^3$ be the vector of probabilities for each cultural sector. The probabilities for each of the 30 classes are then calculated as $P = P_r P_c$. Before computing the probability matrix P, we set a threshold to filter out columns and rows with low probabilities and redistribute these values evenly among the remaining classes.

To obtain vectors P_c and P_r, we trained a Random Forest classifier using the *scikit-learn* Python library on keywords extracted from the documents [28][29] and compared the results with labels assigned by human experts. To avoid overfitting, both models were evaluated using 10-fold cross-validation.

Tables I and II show F1 results obtained using Random Forest models trained to predict cultural sectors, and societal impact domains respectively. Results are presented as macro F1 scores averaged over 10 folds. The obtained classification probabilities are used for the heat map visualized in Figure 1.

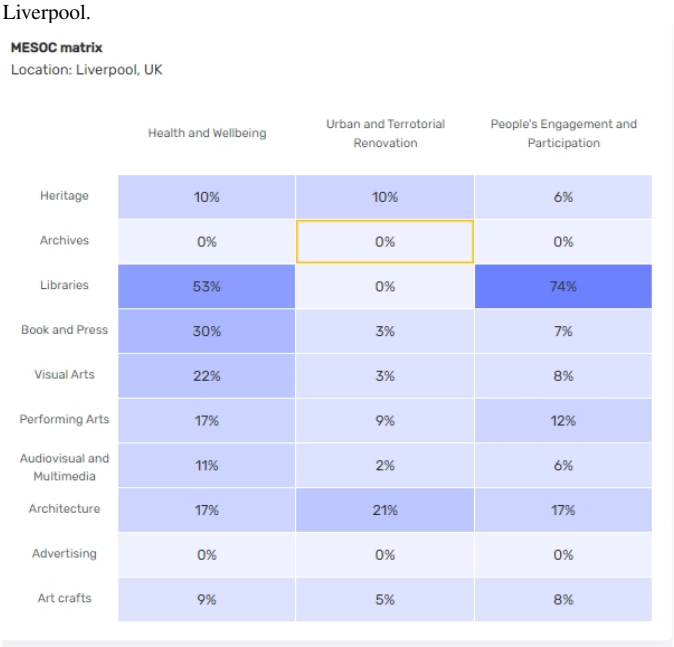
Cultural sector	F1 score
Heritage	0.97
Archives	0.97
Libraries	0.93
Books and Press	0.93
Visual Arts	0.95
Performing Arts	0.93
Audiovisual and Multimedia	0.94
Architecture	0.95
Advertising	0.93
Art crafts	0.97

TABLE I
CLASSIFICATION RESULTS IN TERMS OF MACRO AVERAGED F1 SCORES FOR 10 CULTURAL SECTORS.

Societal impact domain	F1 score
Health and Well-being	0.96
Urban and Territorial Renovation	0.93
People's Engagement and Participation	0.95

TABLE II
CLASSIFICATION RESULTS IN TERMS OF MACRO AVERAGED F1 SCORES FOR 3 SOCIETAL IMPACT DOMAINS.

Fig. 1. Heat map - Mesoc Matrix: classified documents for the city of Liverpool.



C. Similarity Search

Similarity between documents is calculated as Jaccard similarity [11] between a set of keywords from one document and a set of aggregated keywords from every other document that has the same impact or societal domain and cultural sector. Note that the keywords are determined as described in III-A. Let $K_{r,c}$ be the set of keywords from document d describing cultural sector r and impact domain c, and K be the set of aggregated keywords from every other document describing the same domain and cultural sector. Similarity is thus defined as:

$$J(K, K_{r,c}) = \frac{|K \cap K_{r,c}|}{|K \cup K_{r,c}|}. \quad (3)$$

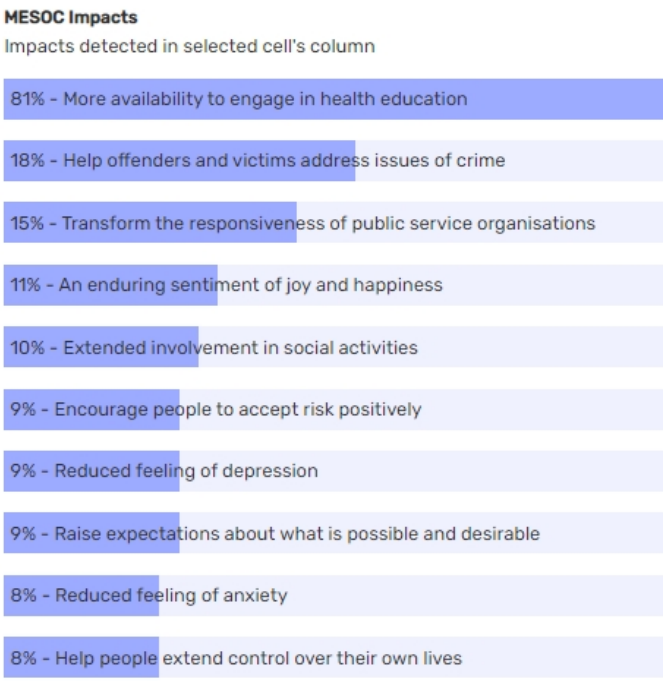
If we are interested in similarity according to a particular impact i, instead of using set $K_{r,c}$, we would aggregate the keywords from each document that has the same impact i. In both cases, we exclude documents that are georeferenced to the same location. The integration of similarity search into Tookit is visualized in Figure 4.

D. Impact Generation

At this final stage of the NLP pipeline, we apply our impact generation and semantic expansion method defined in [14]. The method starts with the initial definition of 100 possible societal impacts defined by a domain expert, and expands their semantic neighborhoods using word2vec embeddings [30], cosine similarity, and K-means clustering. For each impact, a single vector representation was computed by averaging the individual word vector representations captured by the impact. After all bigrams and trigrams are extracted from the dataset, a single vector representation is computed for each of these

n-grams by averaging the single word vector representations associated with each n-gram. Next, the method creates a list of similar extracted n-grams for each impact by calculating semantic similarity using a cosine similarity measure with a threshold of 0.8. To determine the groups of similar impacts, with the aim of avoiding overlaps between their corresponding n-grams, the method uses K-means clustering. After determining the clusters and the final list of impacts with their corresponding most similar n-grams (i.e. the semantic expansion of the impact), the method can automatically annotate new documents. 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. The example of the detected impacts is visualized in Figure 2.

Fig. 2. Impact on the well-being for the studies reported on the city of Liverpool.



IV. Implementation

Next we provide a brief overview of the dataset used and the architecture of the toolkit implementation in the form of a responsive web application.

A. Dataset

The document corpus (i.e., full texts of scientific papers) on the topic of arts and culture contains 573 open-access documents retrieved from academic databases.

Documents are selected by 10 cultural areas according to the UNESCO-EUROSTAT classification [20], [19] and stored in a repository for further experimentation. The length of all 573 documents ranged from 859 characters or 172 word tokens to 696,182 characters or 124,976 word tokens, with an average of 49,894 characters and 8,936 word tokens per document. The dataset is very unbalanced in terms of cultural sectors and societal impact domains, i.e., the classes used in the multi-label classification. To solve this problem, the minority classes were oversampled using the SMOTE technique [31]. Before we proceed with the generation and semantic expansion of impacts, the documents are preprocessed. The first step of preprocessing is the conversion from PDF to text format [32]. PDF conversion produces a lot of noise and unusable characters due to the nature and structure of PDF documents. The problem is solved by removing common abbreviations, hyperlinks, email addresses, phone numbers and unprintable characters, while also cleaning up errors with hyphens, different quotation marks and multiple whitespaces. The second phase of preprocessing involved tokenization of documents, removal of stop words, and normalization. The list of stop words that were removed from the texts came from the Python Natural Language Toolkit (NLTK) [33]. The final step of preprocessing involved lemmatization of the tokens. Lemmatization takes into account the morphological analysis of the tokens and provides a word lemma or base form for all inflected forms of the word. For this last step, we use WordNet Lemmatizer [34] from the NLTK library.

B. Architecture

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 the NLP pipeline (as explained in the Section III), 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.

C. Results and Discussion

The MESOC Toolkit is a georeferenced visualization tool for the semantic search of documents in the MESOC domain of arts and culture and the analysis of their impact on the societal domains. The Figure 3 shows the main screen with georeferenced academic studies pinned on the map of Europe. The map is the entry point for the user inquiry. After selecting the desired location, the user is redirected to the analytical screen with the central visualization element - the heatmap in the form of the MESOC 10x3 matrix shown in Figure 1. The heatmap is the result of the automatic multi-label classification as part of the NLP pipeline, where classification probabilities are used to visualize and color the cells of the heatmap. From the example presented for the city of Liverpool, it can be seen

that documents are linked to libraries and health and well-being by 53% and visual art by 22%.

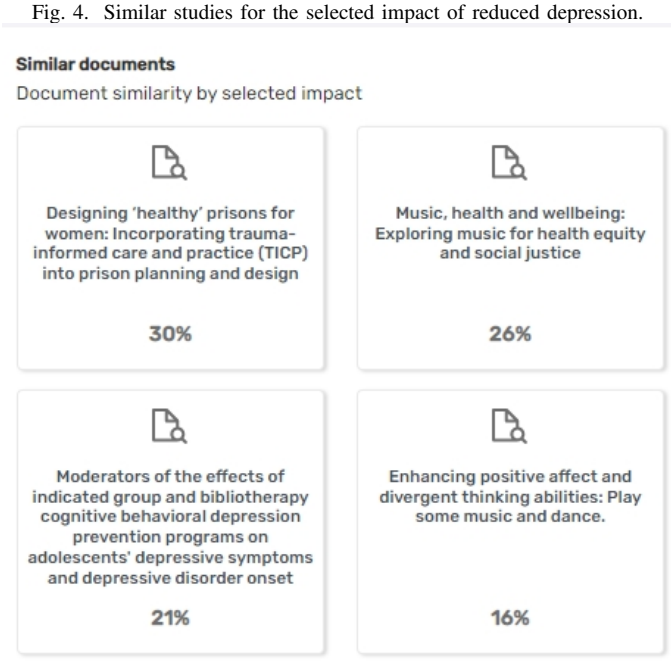
Fig. 3. Map with geo-located academic studies.

After selecting the health and well-being column in the heatmap, the list of detected impacts is displayed in Figure 2. The ”strength” of the detected impact (e.g., ”More availability to engage in health education” - 81%) is used to rank the identified impacts. The toolkit also allows for an explanation of how the impact was derived by listing the semantic expansion used to generate the impact for the document.

Finally, the user can retrieve documents that are similar either in content or in terms of derived impact, as shown in Figure 4.

Integrating NLP methods into BI tools is still an open research issue. Lim et. all in [3] discuss the integration of Big Data Analytics, Text Analytics [4], and Social Networks Analytics [2] as the major research directions for the advancement of business intelligence. In particular, semantic search engines, information extraction, and sentiment analysis [5] are the focus of the underlying NLP methods. In this work, we propose a solution for semantic search by suggesting similar documents either in content or in detected impacts. Moreover, we employ keyword extraction as an information extraction method to create compact document representations that are used to classify documents into 30 different classes, which is inherently a demanding problem. To this end, the research conducted contributes to the integration of NLP methods into a working BI solution in a form of web application.

Fig. 4. Similar studies for the selected impact of reduced depression.



V. Conclusion

In this paper we present a MESOC Toolkit - a georeferenced visualization BI (busines intelligence) tool based on natural language processing (NLP) methods. The main NLP methods in the underlying pipeline are keyword extraction, multi-label text classification, and automatic detection of potential social impacts of cultural policies and practices. The integrated NLP methods were trained with texts from open-access academic publications in the domain of arts and culture collected as part of the Measuring the Social Dimension of Culture (MESOC) project.

The MESOC Toolkit enables exploring of the influence of identified impacts on social value creation in the areas of health and well-being, urban regeneration, and social cohesion. The toolkit also enables semantic search for related documents based on the content of the documents and the identified impacts. Therefore, the research presented can serve as a prototype for measuring societal value by identifying recurrent pathways of transformational processes in society that extend beyond the selected domain of arts and culture.

In the future, we plan to expand the toolkit functionalities by adding more explanations on how impacts are derived; the detection of variables that are indicators of possible pathways of transformation processes in society; the collection of user feedback regarding the correctness and usefulness of automatically generated content and generated analytical insights that can be utilized to facilitate the retraining of machine-learned models and their evaluation.

VI. Acknowledgement

This work was supported by the University of Rijeka under project number uniri-drustv-18-20 and the H2020 project Measuring the Social Dimension of Culture-MESOC. The authors express their gratitude to Erik Jermaniš and Valentin Kuharić who contributed the frontend part of the MESOC Toolkit web application. We also thank Francesco Mollinari for the definition of impacts and for contributions in the explanations of the domain-knowledge.

References

[1] R. Sharda, D. Delen, E. Turban, J. Aronson and T. Liang, Business intelligence and analytics. System for Decision Support.2014.

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

[3] E.-P. Lim, H. Chen and G. Chen, "Business intelligence and analytics: Research directions." ACM Transactions on Management Information Systems (TMIS) 3.4: 1-10. 2013

[4] W. Chung and T. L. B. Tseng, "Discovering business intelligence from online product reviews: A rule-induction framework." Expert systems with applications 39.15: 11870-11879.2012.

[5] I. Sreesurya, H. Rathi, P. Jain and T. K. Jain, Hypex: A Tool for Extracting Business Intelligence from Sentiment Analysis using Enhanced LSTM. Multimedia Tools and Applications, 79, 35641-35663. 2020.

[6] D. Jurafsky and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall. 2009.

[7] S. Beliga, A. Meštrović and S. Martinčić-Ipšić, "An overview of graph-based keyword extraction methods and approaches." Journal of information and organizational sciences 39.1. 1–20.2015

[8] S. Beliga, A. Meštrović and S. Martinčić-Ipšić, "Selectivity-based keyword extraction method." International Journal on Semantic Web and Information Systems (IJSWIS) 12.3. 1–26.2016.

[9] P. K. Bogović, A. Meštrović, S. Beliga and S. Martinčić-Ipšić, 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.

[10] S. Martinčić-Ipšić, T. Miličić and L. Todorovski, The Influence of Feature Representation of Text on the Performance of Document Classification. Applied Science. 9, 743. 2019.

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

[12] K. Babić, M. Petrović, S. Beliga, S. Martinčić-Ipšić, M. Matešić and A. Meštrović, Characterisation of COVID-19-related tweets in the Croatian language: framework based on the Cro-CoV-cseBERT model. Applied Sciences, 11(21), 10442.2021.

[13] S. Beliga, S. Martinčić-Ipšić, M. Matešić, I. P. Vuksanović and A. Meštrović, 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.

[14] P. K. Bogović, F. Molinari, B. Kovačić and S. Martinčić-Ipšić, Generation and Semantic Expansion of Impacts in Arts and Culture, FICC, Data Science, San Francisco, USA, 2022. (accepted)

[15] MESOC: Measuring the social dimension of culture. https://www.mesoc-project.eu/

[16] European Commission: Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. A New European Agenda for Culture. COM/2018/267 final (2018)

[17] Council of the European Union: Council conclusions on cultural and creative crossovers to stimulate innovation, economic sustainability and social inclusion. Official Journal of the European Union, C 172, 27 May 2015 (2015)

[18] Council of Europe: Indicator Framework on Culture and Democracy (IFCD) (2013) https://www.coe.int/en/web/culture-and-heritage/indicators-culture-and-democracy

[19] 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

[20] EUROSTAT: The ESS-Net Culture Framework (2012). https://ec.europa.eu/assets/eac/culture/library/reports/ess-net-report_en.pdf

[21] F. Matarasso, Use or Ornament? The social impact of participation in the arts. Comedia. 1997

[22] European Parliament and Council: Decision No 445/2014/EU of 16 April 2014 establishing a Union action for the European Capitals of Culture for the years 2020 to 2033 and repealing Decision No 1622/2006/EC. Official Journal of the European Union, L 132, 3 May 2014 (2014)

[23] F. Boudin, "pke: an open source python-based keyphrase extraction toolkit." Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations 69–73 2016.

[24] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features." Information Sciences 509 257-289.2020.

[25] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text." Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing 404–411 2004.

[26] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin and C. G. Nevill-Manning, "KEA: practical automatic keyphrase extraction." In Proceedings of the fourth ACM conference on Digital libraries (DL '99). Association for Computing Machinery, New York, NY, USA, 254–255. 1999. DOI:https://doi.org/10.1145/313238.313437

[27] C.D. Manning, P. Raghavan and H. Schütze, "Introduction to Information Retrieval"; Cambridge University Press: New York, NY, USA, 2008.

[28] L. Breiman, Random forests. Machine learning 45.1: 5-32. 2001.

[29] L. Buitinck, et al, API design for machine learning software: Experiences from the scikit-learn project. In Proceedings of the ECML PKDD Workshop: Languages for Data Mining and Machine Learning, Prague, Czech Republic, 23–27 September. 108–122. 2013.

[30] K. Babić, S. Martinčić-Ipšić and A. Meštrović, Survey of Neural Text Representation Models. Information 2020, 11, 511. 2020. https://doi.org/10.3390/info11110511

[31] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique" Journal of Artificial Intelligence Research 16 321–357. 2002. DOI:http://dx.doi.org/10.1613/jair.953

[32] J. Mckie, PyMuPDF - the Python bindings for MuPDF (2021). https://buildmedia.readthedocs.org/media/pdf/pymupdf/latest/pymupdf.pdf

[33] S. Bird, E. Klein and E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.", 2009.https://www.nltk.org/

[34] C. Fellbaum, WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press (1998, ed.). https://doi.org/10.7551/mitpress/7287.001.0001