This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870935

Research Classs:

InfoCov & MESOC Joint Research Workshop

Date: Friday June 11th 2021 (9 AM – 5 PM)

Venue: Department of informatics, University of Rijeka, Radmile Matejčić 2, Classroom O-028

Online: https://meet.jit.si/infocov-mesoc-workshop

Program Committee:

Sanda Martinčić-Ipšić, InfoCov & MESOC project

Ana Meštović, InfoCov project

Božidar Kovačić, MESOC project

InfoCoV Schedule

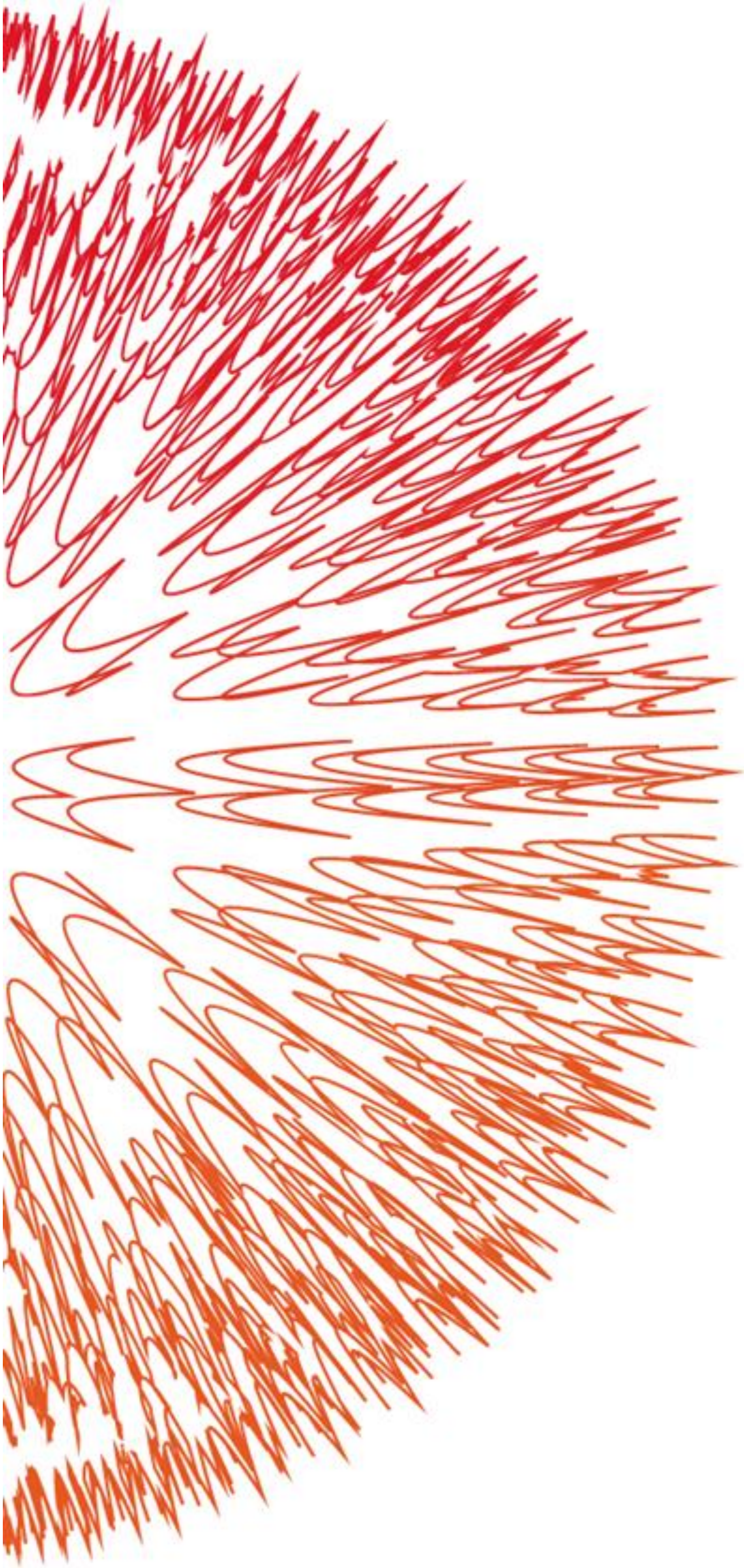
https://infocov.uniri.hr/

9:00 – 9:45	Ana Meštrović: An Overview of the InfoCoV Project Research
9:45 – 10:30	Slobodan Beliga: Information Monitoring of Croatian COVID-19 Related Online News During the Pandemic in 2020
10:30 – 11:00	**Coffee break**
11:00 – 11:45	Karlo Babić: Twitter Analysis with Machine Learning
11:45 – 12:30	Milan Petrović: Twitter Scraping and Data Processing
12:30 – 13:30	**Lunch break**

MESOC Schedule

https:/mesoc-project.eu/

13:30 – 13:55	Božidar Kovačić: MESOC Repository of Documents
13:55 – 14:20	Sanda Martinčić-Ipšić: MESOC Toolkit – NLP Functionalities and Visualizations
14:20 – 14:45	Francesco Molinari: From Textual to Predictive Analytics for Impact Assessment: An Experiment in Ontology-based Search for Measurable patterns
14:45 – 15:00	Dragan Čišić: MESOC SERAPEUM
15:00 – 15:15	**Coffee break**
15:15 – 15:45	Petar Kristijan Bogović: Detection and Semantic Expansion of Impacts from Documents in the Domain of Culture
15:45 – 16:00	Dino Aljević: Document Classification into the MESOC Matrix
16:00 – 16:15	Erik Jermaniš: Development of MESOC Toolkit Web Application in React
16:15 – 16:30	Valentin Kuharić: MESOC Client-side Web Application Architecture in Node.js, Express.js and React
16:30 – 16:45	Dino Aljević: MESOC Tookit Backend
16:45 – 17:00	Sanda Martinčić-Ipšić: MESOC Toolikt Demo



InfoCoV

Title: An Overview of the InfoCoV Project Research

Speaker: **Associate professor Ana Meštrović, PhD,** University of Rijeka, Department of Informatics and Center for Artificial Intelligence and Cybersecurity

Abstract: Project "*Multilayer Framework for the Information Spreading Characterization in Social Media during the COVID-19 Crisis - InfoCoV*" funded by the Croatian science foundation is focused on the research of the social aspects of COVID-19 pandemic.

Social media accelerate the information spreading and may cause an infodemic, especially during the crisis. As stated by the WHO, the COVID-19 outbreak culminated with a massive infodemic, which is potentially dangerous because it makes difficult for individuals to find reliable sources of information when they need it. Furthermore, COVID-19 related communication crisis brings new challenges in terms of large communication volumes, massive datasets, new terminology, new aspects and new specific topics that have come into focus. The aim of this research is to help with a better understanding of crisis communication by using NLP methods and techniques. Additionally, we combine social network analysis (SNA) methods.

For the purpose of this research, we collect and analyse empirical data crawled from social networks (Twitter, Reddit, Youtube, etc.) and Croatian online portals. Next, we perform a quantitative analysis of texts and messages in media published during the COVID-19 pandemic in terms of automatic detection of key terms, named-entity recognition, topic modelling and automatic classification of positive, neutral and negative attitudes, etc. We track the dynamics of changing the communication trends in social media during the pandemic.

Title: Information Monitoring of Croatian COVID-19 Related Online News During the Pandemic in 2020

Speaker: **Slobodan Beliga, PhD,** University of Rijeka, Department of Informatics and Center for Artificial Intelligence and Cybersecurity

Abstract: The presentation will introduce the problems and challenges of information monitoring in the Croatian media space related to the situation caused by the appearance of the SARS-CoV-2 virus, specifically in online newspaper publications. The results of the longitudinal study will be presented through the perspective of natural language analysis of online news that was published on Croatian mainstream newspaper portals. The study analyzes the frequency of writing about coronavirus in the 1st and 2nd epidemic wave. Moreover, an investigation on the possible association between infodemia and the influence of the environment (i.e. real-world factors), which we believe could strongly influence the representation and amount of COVID-19-related news, will be presented. For example, the number of infected people or the number of deaths caused by COVID-19 disease. This is followed by the analysis of the news articles' content in terms of the most frequently used COVID-19 news-specific terminology, and tracking changes across 13 months of a pandemic. In addition, the main actors who were most represented in the media during the pandemic year were identified, in terms of those who most often gave statements to the media or those that journalists frequently write about. Institutions and organizations that were most often mentioned in the news during pandemics were also monitored. Based on the results of a one-year study, the presentation will enable a better understanding of crisis communication and provide some guidelines for future monitoring of coronavirus-induced infodemia.

Title: Twitter Analysis with Machine Learning

Speaker: **Karlo Babić,** University of Rijeka, Department of Informatics and Center for Artificial Intelligence and Cybersecurity

Abstract: First, we analyzed and compared Croatian and Polish Twitter datasets. After collecting tweets related to COVID-19 in the period from 20.01.2020 until 01.07.2020, we automatically annotated positive, negative, and neutral tweets with a simple method, and then used a classifier to annotate the dataset again. To interpret the data, the total number as well as the numbers of positive and negative tweets are plotted through time for Croatian and Polish tweets. Second, we explored the influence of COVID-19 related content in tweets on their spreadability. The experiment is performed in two steps on the dataset of tweets in the Croatian language posted during the COVID-19 pandemics. In the first step, we train a feedforward neural network model to predict if a tweet is highly-spreadable or not. In the second step, we use this model in a set of experiments to predict the average spreadability of tweets. Third, we are fine-tuning a BERT model, which is trained on Slavic languages, on Croatian texts related to COVID-19. The goal is to have a well-performing model that we can use for NLP tasks such as sentiment analysis on Croatian tweets related to COVID-19. We will compare different fine-tuning methods: fine-tuning on Croatian texts related to COVID-19, fine-tuning on a task (sentiment analysis), fine-tuning on additional Croatian Twitter datasets, etc.

Title: Twitter Scraping and Data Processing

Speaker: **Milan Petrović,** University of Rijeka, Department of Informatics and Center for Artificial Intelligence and Cybersecurity

Abstract: In this work it is presented the processes related to the collection of data from Twitter. Data collection, cleaning and analysis procedures are explained. Data were collected specifically for the Croatian region, i.e. the content in the Croatian language was analyzed. It shows the use of the Twitter API and its capabilities, as well as an overview of the obtained metadata. From the collected data, those are selected that can be used to monitor the information spreading. The analysis was performed using complex networks. In addition, a dataset was created from the collected data that can be used for machine learning and natural language processing.

MESOC

Title: MESOC Repository of documents

Speaker: **Božidar Kovačić,** PhD, University of Rijeka, Department of Informatics

Abstract: MESOC Repository of documents is searchable collection of thematic publications of relevant cultural policies and practices to extract the most appropriate impact transmission variables and indicators in retrospect, and to analyze what have been the critical success factors in determining the final outcomes of the selected transition pathway. Their successful deployment will be associated to the academic profile of the research team coordinating the efforts purporting to their delivery. Activity consists in collecting, selecting and analysing relevant documents on the social impacts of cultural policies, covering all the three dimensions of Health and Well Being, Urban and Territorial Renovation, People's Engagement and Participation. Repository offers features for organizing documents in different collections and it will be available on the official project website, even after the project's end. Technology for development of Online Document Repository is based on MVC model and includes Relational database MySQL, PHP programing language and Phalcon framework. Development of communication module for data exchange between the repository of document and the MESOC Toolkit web application enables fast access and retrieval of data from the document repository and the MESOC Toolkit app.

Title: MESOC Toolkit – NLP Functionalities and Visualizations

Speaker: **Sanda Martinčić-Ipšić,** PhD, University of Rijeka, Department of Informatics and Center for Artificial Intelligence and Cybersecurity

Abstract: MESOC Toolkit is a natural language processing powered analytical tool aimed to discover the impacts of cultural policies and practices on society. Toolkit is tasked with the automatic discovery of hidden semantic structures under the perspective of societal value creation and uncovering latent transmission processes. MESOC toolkit is a georeferenced visualization dashboard system organized around MESOC matrix. The MESOC matrix cross-references cultural domains (introduced by the EUROSTAT) against the three pillars of the structural model. Ten cultural domains are Heritage, Archives, Libraries, Book and Press, Visual Arts, Performing Arts, Audiovisual and Multimedia, Architecture, Advertising and Art crafts. Three pillars of the structural model stem from three crossover themes of the new European Agenda for Culture: Health and Wellbeing, Urban and Territorial Renovation and Social Cohesion. Underlaying NLP engine incorporates multiclass document classification, the discovery of impacts and semantic retrieval according to the analyzed content. The toolkit is implemented as the web app capable of two use case scenarios: (1) georeferenced explorative analysis of city use cases (red) and scientific (blue) studies and (2) automatic analysis of user-generated content.

Title: From Textual to Predictive Analytics for Impact Assessment: An Experiment in Ontology-based Search for Measurable patterns

Speaker: **Francesco Molinari**, MESOC project manager

Abstract: Knowledge discovery in databases is often referred to as the ultimate goal of text mining (Hotho et al., 2005). Unfortunately, there is usually a gap between information retrieval (however accurate and reliable) and the generation of new, relevant and interesting interpretations and implications from the lexical analysis carried out. This gap, at least to some extent, resembles to and derives from the familiar hiatus between data collection and pattern visualisation through "connecting the dots" (Moioli, 2020). But there is more, which is probably related to an imperfect understanding of the context which most of the written resources analysed belong to. Such issue calls to an extra effort even the so-called domain experts, who all-too-often indulge in a generic demand for "explainable NLP" putting an extra burden on the AI technologists who are involved in the heuristic trials. In this framework, the H2020 funded MESOC project is now carrying out an innovative experiment of lexical analysis and predictive classification of academic papers describing the societal impacts of cultural policies and activities. The experiment relies on a combination of state-of-the-art text mining methods and tools with an extended ontology of the cultural domain and a structural model describing the possible directions of impact generation. The aim is to identify the most frequently recurring patterns and propose alternative ways of measuring them by appropriate statistical proxies. This should contribute to improving domain expert knowledge on societal impact mechanisms and the key actors and processes related to them.

References:

Hotho, A., Nürnberger, A. and Paaß, G. (2005). "A brief survey of text mining". In Ldv Forum, Vol. 20(1), pp. 19-62. Online: https://www.semanticscholar.org/paper/A-Brief-Survey-of-Text-Mining-HothoN%C3%BCrnberger/8f74f5623c4e5c5931641a264cfd7c02097e1e22

Moioli, F. (2020). Data is only the beginning. Connecting the dots creates Knowledge. However, it is Wisdom which translates insights into impactful decisions. LinkedIn Pulse, 8 January. Online: https://www.linkedin.com/pulse/data-only-beginning-connecting-dots-creates-knowledge-fabio-moioli

Title: MESOC SERAPEUM

Speaker: **Dragan Čišić**, PhD, University of Rijeka, Department of Informatics

Abstract: MESOC Serapeum is an AI playground for the MESOC partners. The idea is to create a system that would introduce consortium members into AI. The system, therefore, uses several different methods for NLP, especially transformers. Facebook's FastAi, Google's BERT, and OpenAIi GPT2 are especially represented. The system contains 7 different models of supervised learning and 2 models of transfer learning. The WEB site is currently hosted by the University of Valencia.

Title: Detection and Semantic Expansion of Impacts from Documents in the Domain of Culture

Speaker: **Petar Kristijan Bogović**, University of Rijeka, Department of Informatics

Abstract: The dataset used for this research was provided by the MESOC toolkit and uploaded by users. The documents were originally in PDF format, so we needed to convert them first to a textual format to apply different NLP procedures. Following the conversion, the references sections were removed to reduce unnecessary noise in the data. On this data, we first applied an unsupervised Named Entity Recognition procedure to automatically extract mentioned locations in the text and processed them using a geocoding Python library to find their respective coordinates (LAT, LON). Secondly, we applied an unsupervised automatic keyword extraction method YAKE to extract each documents corresponding keywords, so they could be used to classify each document into one or multiple cells of the MESOC matrix. Lastly, we focused on the automatic extraction of impacts. The impact can be defined as a change expected to happen due to the implementation and application of a given policy or event, with a focus on the health and wellbeing of citizens, urban and territorial renovation, and people’s engagement and participation. After defining a list of different impacts, we converted these impacts to word2vec embeddings and clustered them to identify impacts that overlapped together the most. To visualize these clusters, we used t-SNE. Next, we focused on expanding the semantic space of each impact by extracting all the bigrams and trigrams from the documents and converting these ngrams to word2vec embeddings. These ngram word2vec embeddings were then compared to word2vec embeddings of impacts using cosine similarity to define a list of ngrams semantically most similar to each impact. After defining these lists, automatic detection of these impacts was developed by scanning each document for ngrams defined in the lists corresponding to each impact.

Title: Document Classification into the MESOC Matrix

Speaker: **Dino Aljević,** University of Rijeka, Department of Informatics

Abstract: The precursor to semantic retrieval of related documents is a classification of uploaded documents into 30 different classes which correspond to rows and columns of the MESOC matrix. An already complex task of multi-label classification is made more complex by low-quality data and the unclear distinction between different domains described by the matrix. Our somewhat unorthodox method attempts to solve this problem by combining two different Random Forest models trained to classify document into columns and rows respectively. In this presentation, we will take a detailed look at the inner workings of this method, performance and challenges we faced along the way.

Title: Development of MESOC Toolkit Web Application in React

Speaker: Erik Jermaniš, University of Rijeka, Department of Informatics

Abstract: The development process of MESOC toolkit web application consists of two parts. The first part is creating the application. This part includes designing the UI and UX, creating mock-ups and then transforming them into React code. The second part is focused on connecting the application with the API and making everything work.

Title: MESOC Client-side Web Application Architecture in Node.js, Express.js and React

Speaker: Valentin Kuharić, University of Rijeka, Department of Informatics

Abstract: The presentation describes the requirements and the thought-process of designing the architecture for the frontend part of the web-based application MESOC Toolkit. It also includes the process required to set up and deploy the application. The user management process will also be described in the react application. Application is written in Javascript using the Node.js runtime environment, enabling us to run Javascript code outside of the browser. The application consists of two parts: The Nodejs + Express.js based server and the React-based single-page application. This design pattern enables us to have the development workflow and advantages of SPA while having the SEO advantages for the homepage.

Title: MESOC Toolkit Backend

Speaker: **Dino Aljević**, University of Rijeka, Department of Informatics

Abstract: The backend is an integral part of the MESOC Toolkit containing core functionality, business logic, and NLP engine hidden behind a RESTful API. In this talk, we overview Toolkit's architecture, requirements and constraints which shaped the backend design. Finally, we outline the document processing NLP pipeline and principles of incorporation of machine learning methods to facilitate the analysis of user-uploaded documents.