KU LEUVEN

Faculty of Science

Title

Group 8

Keith Atienza – r1015961
Michael Carlo – r0863809
Fabrizio Sacco – r0876560
Alexander Thomas Savvides – r0873882

Modern Data Analytics, Project Proposal

Academic year: 2024-2025

Professor Dr. Martial Luyts

FACULTY OF SCIENCE CAMPUS LEUVEN

CELESTIJNENLAAN 200

Introduction

The Horizon Europe program, with its €95.5 billion budget for 2021–2027, is one of the largest funding programs globally, aimed at fostering research and innovation across the European Union. It covers critical areas such as climate change, digitalization, health, and security. However, the complexity of the program and the sheer volume of projects funded make it difficult for stakeholders, including policymakers, researchers, and citizens, to efficiently navigate and understand the initiatives and their outcomes. In this context, our project aims to address this challenge by developing a comprehensive data-driven platform that analyses the funded projects under Horizon Europe, providing users with insights into research themes, funding distribution, and collaboration networks.

Our approach leverages specifically Natural Language Processing (NLP), unsupervised learning, and various machine learning algorithms to analyse the CORDIS dataset, which contains detailed metadata on Horizon Europe projects. This analysis will be instrumental in revealing trends in research priorities, identifying gaps in funding, and facilitating collaboration among project participants. We will focus on the application of topic modelling, NLP techniques for semantic analysis, and a Retrieval-Augmented Generation Language Model (RAG-LLM) for summarization and querying, to provide a transparent, accessible platform for stakeholders to explore Horizon Europe's vast landscape.

Approach

The objective of this project is to analyse the research priorities, funding allocations, and collaborative dynamics within the Horizon Europe framework. The approach that will follow now relates established methods in unsupervised learning, text analysis, and network theory to uncover the underlying trends and patterns in Horizon Europe projects.

Data Processing and Topic Modelling

The first step in our analysis involves data preprocessing. The CORDIS dataset contains a wide range of information, including project titles, descriptions, funding amounts, participant organizations, and research themes. This data must be carefully cleaned and structured before any meaningful analysis can take place. We will handle missing data, normalize textual descriptions, and standardize categorical variables such as funding mechanisms, participant roles, and program classifications.

Once the data is prepared, we will use **topic modelling** to identify the primary research themes present across Horizon Europe projects. Topic modelling is a technique from unsupervised learning that allows us to extract latent structures from textual data. The most commonly used method for this purpose is **Latent Dirichlet Allocation (LDA)**, which will be employed to extract thematic clusters from the project descriptions and summaries. LDA assumes that each project description is a mixture of various topics, and the goal is to infer the distribution of these topics across the entire dataset. This technique will allow us to identify which research themes are predominant in the Horizon Europe projects and how these themes have evolved over time.

Additionally, we will use **BERT (Bidirectional Encoder Representations from Transformers)**, an advanced model for natural language processing and understanding, to enhance the overall quality of the extracted topics. BERT's ability to capture contextual information in a text will help us identify subtler, less visible themes that might be missed by traditional topic modeling methods. Using BERT's embeddings, we will cluster project descriptions into related groups, providing deeper insights into the relationships between research areas.

Exploring Research Trends and Gaps

One of the key objectives of Horizon Europe is to address global challenges, including climate change, health, digitalization, and security. By analysing the distribution of funding across research themes, we aim to answer several critical questions: Which topics receive the most funding from the EU? Are there research areas that are underfunded relative to their global importance? Furthermore, we will assess whether the projects funded under Horizon Europe align with the overarching goals of the initiative, particularly in terms of promoting EU competitiveness and addressing global challenges.

Implementation of a Retrieval-Augmented Generation (RAG) Model

To make the findings more accessible to a wide audience, we will implement a **Retrieval-Augmented Generation (RAG)** model, integrated with a **Large Language Model (LLM)** such as GPT, to generate clear, concise summaries of the research projects and their outcomes. The RAG-LLM system will allow users to query the dataset in natural language and receive real-time, data-driven responses. This will enable policymakers, researchers, and the general public to quickly access relevant information about Horizon Europe projects, including funding distribution, research themes, and collaborative networks.

Data Visualization and Interactive Tools

To complement the RAG-LLM model, we will design an interactive dashboard using **Plotly** and **Dash**, which will allow users to explore the data visually. The dashboard will provide several interactive features, including:

1. **Thematic Exploration:** Users can explore the various research themes funded by Horizon Europe, with the ability to drill down into specific areas such as climate change, health, and digitalization.
2. **Funding Distribution:** The dashboard will display funding allocations across different research areas, highlighting trends and gaps in the funding landscape.
3. **Project Summaries:** Users can view detailed summaries of individual projects, including their objectives, outcomes, and funding levels.

These interactive tools will enhance the transparency and accessibility of the data, enabling stakeholders to better understand the impact of Horizon Europe on research and innovation.

End Product and Impact

The end product will be a comprehensive, interactive platform that provides insights into the priorities, funding allocations, and collaborative dynamics of Horizon Europe. This tool will serve as a valuable resource for policymakers, researchers, and the general public, offering a clear and actionable understanding of how EU funds are being allocated to address global challenges and enhance European competitiveness. By applying the above mentioned methods, we aim to bring transparency and clarity to the Horizon Europe framework, helping stakeholders make more informed decisions about future research initiatives.

Through this project, we will provide a better understanding of how Horizon Europe is shaping the future of science, technology, and innovation in Europe, and offer recommendations for improving the program's focus and impact. Ultimately, our work will contribute to a more informed and efficient European Research Area, facilitating greater collaboration and innovation across the continent.