

Predicting the Tendency Toward Open Science in Flemish Research Projects

Hoang-Son Pham^{1,2} and Amr Ali-Eldin^{1,2}

¹ Centre for Research & Development Monitoring (ECOOM-UHasselt), ² Data Science Institute, Hasselt University, Hasselt, Belgium.

Presented at Open Science Network Day 2025
22nd of May 2025, Brussels

INTRODUCTION

This paper presents a work in progress and a novel machine learning-based approach to assess and predict the tendency to support open-access within research projects as a feature of Open Science support.

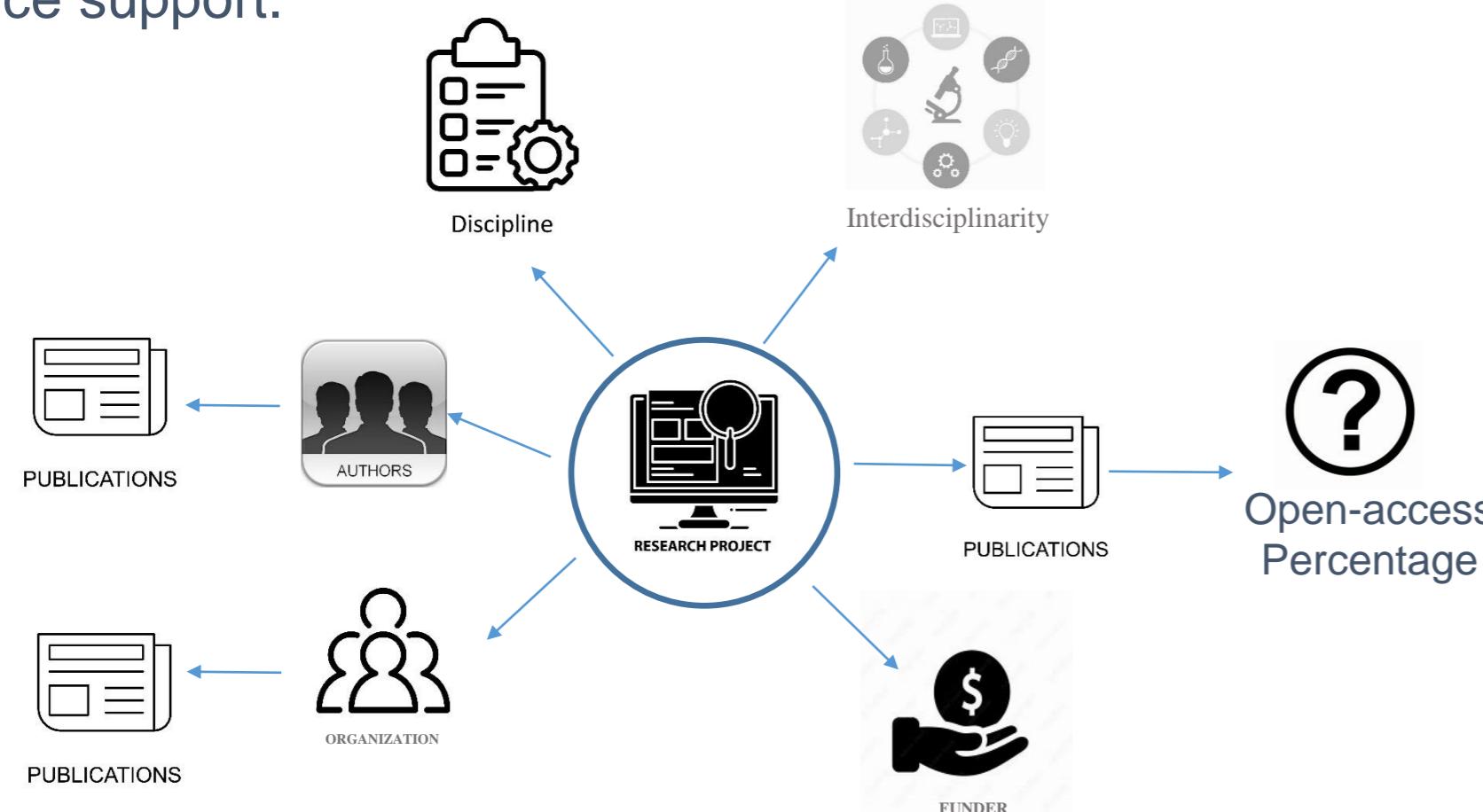


Fig.1. Project related information

In this work, the tendency to OA support of a project is represented by the percentage of open-access articles expected as outcome of that project.

OA support is expected to be influenced by various project-related factors, such as open-access journal articles authored by project participants (researchers or organizations), the funding source, the associated research disciplines, and the interdisciplinarity of the project.

APPROACHES

Analyzing key indicators such as publication practices, funding sources, research disciplines, and interdisciplinarity, we develop predictive models that identify open-access support level.

Approach 1. Apply Predictive Machine Learning

Step 1: Develop feature creation algorithm to generate features for regression models

Algorithm 1 Feature Creation Algorithm

Input: projectID

Output: *F*

1: // Step 1: Count percentage of open-access publications

2: *P* \leftarrow GetPublications(projectID)

3: *p* \leftarrow CalculatePercentageOpenAccess(*P*)

4: // Step 2: Calculate average percentage of open-access publications for researchers and research groups

5: *R* \leftarrow GetResearchers(projectID)

6: *r* \leftarrow CalculateAveragePercentageResearchers(*R*)

7: *O* \leftarrow GetResearchGroups(projectID)

8: *o* \leftarrow CalculateAveragePercentageGroups(*O*)

9: // Step 3: Get Funder

10: *f* \leftarrow GetFunder(projectID)

11: // Step 4: Predict Disciplines

12: *d* \leftarrow PredictDisciplines(projectID)

13: // Step 5: Calculate IDR

14: *DR, DO, DD* \leftarrow CalculateIDR(projectID)

15: // Output the feature set

16: *F* \leftarrow $\{p, r, o, d, f, DR, DO, DD\}$

17: return *F*

Step 2: Train Machine Learning Model to predict open-access support

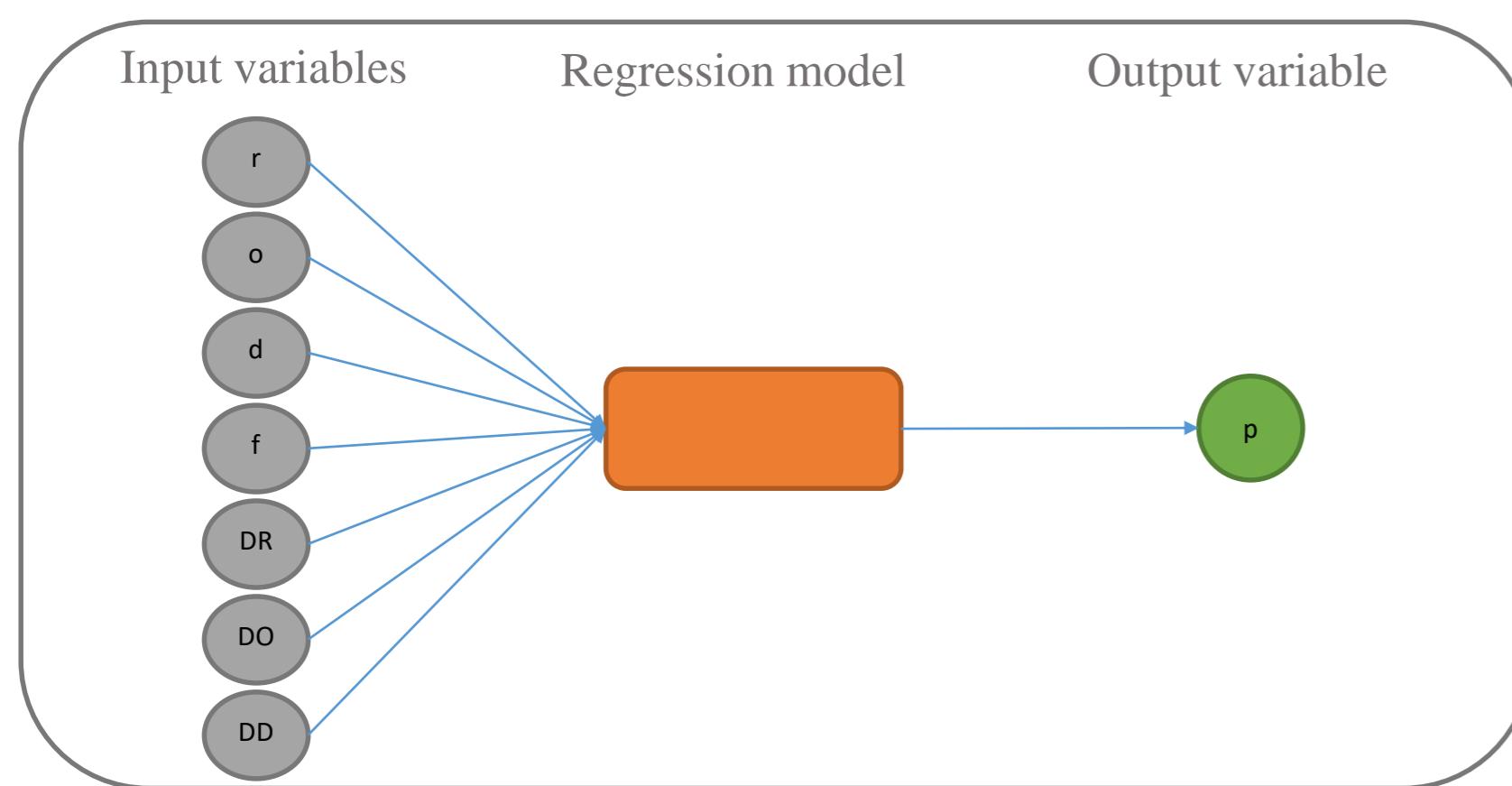


Fig.2. Illustration of regression model

Approach 2. Apply Large Language Model

LLM techniques like ChatGPT and LLaMA were applied to predict OA support of research project.

Given project abstract, disciplines, funder, etc. we asked LLM how likely is it that the resulting publications will be open access?

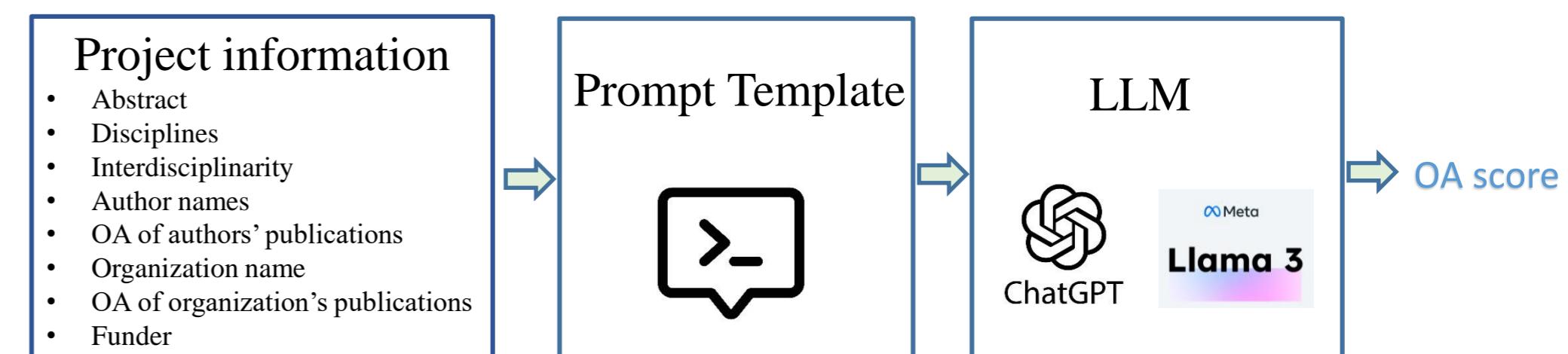


Fig.3. Illustration of LLM model application

RESULTS

The models were trained and tested on 703 research projects from FRIS portal.

ML model performance

	MSE	Precision	Recall	F1-Score	Accuracy
LR	0.074	0.78	0.80	0.80	0.80
RFR	0.074	0.82	0.84	0.82	0.85
SVR	0.074	0.78	0.80	0.79	0.80
NN	0.096	0.83	0.76	0.76	0.76
LSTM	0.074	0.82	0.81	0.82	0.81
BiLSTM	0.087	0.81	0.75	0.78	0.76
CNN	0.083	0.84	0.76	0.79	0.76

LLM results

```
# define prompt template
prompt_template = ("Open-access (OA) support level (value from 0 to 1) of a project is represented by the percentage of publications that were published in open access. " +
"Given the following information of a project:\n" +
"Abstract: {abstract}\n" +
"Disciplines: {d}\n" +
"Diversity of authors which is calculated by using authors' disciplines: {DR}\n" +
"Author names: {authors}\n" +
"Percentage of OA of authors: {r}\n" +
"Organization names: {o}\n" +
"Percentage of OA of organizations: {o}\n" +
"\"Funder: {f}\n" +
"Using all provided information to predict the open-access support level of this project.\n" )
```

PROMPT TEMPLATE

Project 1 : To predict the open-access (OA) support level of the project based on the provided information, we can consider the open-access percentages of the authors and organizations involved, as well as the diversity of authors' disciplines.

- **Author OA Percentage**:
- The percentage of OA for the authors is 0.05 (5%).
- **Organization OA Percentage**:
- The percentage of OA for the organizations is 0.1 (10%).
- **Diversity of Authors**:
- The diversity of authors is calculated at 0.1246. This represents the variety of disciplines (Languages and literary studies, Psychology and cognitive sciences) that might influence publication practices related to open access.

Final Result
- **Open-access support level**: **0.1373** (approximately 14% when expressed as a percentage, acknowledging regional or specific practices in OA that might affect interpretation).

LLama3 Model performance

- Precision: 76%
- Recall: 83%
- F1-score: 78%
- Accuracy: 83%

CONCLUSION

We developed and evaluated machine learning and large language models to predict the open access (OA) support level of research projects. Both approaches demonstrated strong performance, achieving an accuracy of 83% - 85% on our evaluation dataset. These results highlight the potential of AI-driven methods to support open science monitoring and decision-making.

In future work, we aim to enhance model interpretability and generalizability by incorporating more diverse datasets, expanding feature sets (e.g., author OA history, funding policies), and applying the models across different funding agencies and research domains.

Acknowledgment: This study was supported by The Expertise Center for Research and Development Monitoring (ECOOM), Flanders, Belgium.