# An organizational approach for discipline prediction in research projects

**Hoang-Son Pham**

ECOOM UHasselt

EuroCRIS membership meeting, May 31, 2023, Brussels

# Outline



ecoom
EXPERTISECENTRUM O&O MONITORING

UHASSELT

# Introduction

RIS systems store research information: publications, projects

# Introduction

RIS systems store research information: publications, projects

Classifying research documents is an essential task

# Introduction

RIS systems store research information: publications, projects

Classifying research documents is an essential task

Automatically classifying disciplines is challenging

# Introduction

RIS systems store research information: publications, projects

Classifying research documents is an essential task

Automatically classifying disciplines is challenging

Current approaches: citation-based, machine learning algorithms

ecoom
EXPERTISECENTRUM O&O MONITORING

UHASSELT

# Introduction

Project metadata

- title
- keywords
- abstract
- researchers
- organisations
- disciplines

ecoom
EXPERTISECENTRUM O&O MONITORING

UHASSELT

# Introduction

Project metadata

- title
- keywords
- abstract
- researchers
- organisations
- disciplines

Machine learning approach for discipline prediction


ecoom
EXPERTISECENTRUM O&O MONITORING

UHASSELT

# Proposed approach

Predict disciplines of project based on disciplines of researchers

Researcher disciplines

- profile
- organizations
- projects
- co-authors on projects
- publications
- co-authors on publications

ecoom
EXPERTISECENTRUM O&O MONITORING

UHASSELT

# Proposed approach

Each researcher is represented by a matrix (N x 6)

| | profile | organization | projects | co-authors on projects | publications | co-authors on publications |
|---|---|---|---|---|---|---|
| **discipline 1** | 0.1 | 0.2 | 0.3 | 0.1 | 0.1 | 0.2 |
| **discipline 2** | 0.2 | 0.3 | 0 | 0 | 0.2 | 0.1 |
| **discipline 3** | 0.3 | 0.4 | 0.2 | 0.3 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... |
| **...** | ... | ... | ... | ... | ... | ... |
| **discipline N** | ... | ... | ... | ... | ... | ... |

where a cell (i,j) is frequency of discipline i in resource j

ecoom
EXPERTISECENTRUM O&O MONITORING

UHASSELT

# Proposed approach

Each researcher is represented by a matrix (N x 6)

| | profile | organization | projects | co-authors on projects | publications | co-authors on publications |
|---|---|---|---|---|---|---|
| **discipline 1** | 0.1 | 0.2 | 0.3 | 0.1 | 0.1 | 0.2 |
| **discipline 2** | 0.2 | 0.3 | 0 | 0 | 0.2 | 0.1 |
| **discipline 3** | 0.3 | 0.4 | 0.2 | 0.3 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... |
| **...** | ... | ... | ... | ... | ... | ... |
| **discipline N** | ... | ... | ... | ... | ... | ... |

where a cell (i,j) is frequency of discipline i in resource j

A project is represented by a matrix (N x 6), i.e., sum of researchers matrices

ecoom
EXPERTISECENTRUM O&O MONITORING

UHASSELT

# Proposed approach

Each researcher is represented by a matrix (N x 6)

| | profile | organization | projects | co-authors on projects | publications | co-authors on publications |
|---|---|---|---|---|---|---|
| **discipline 1** | 0.1 | 0.2 | 0.3 | 0.1 | 0.1 | 0.2 |
| **discipline 2** | 0.2 | 0.3 | 0 | 0 | 0.2 | 0.1 |
| **discipline 3** | 0.3 | 0.4 | 0.2 | 0.3 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... |
| **...** | ... | ... | ... | ... | ... | ... |
| **discipline N** | ... | ... | ... | ... | ... | ... |

where a cell (i,j) is frequency of discipline i in resource j

A project is represented by a matrix (N x 6), i.e., sum of researchers matrices

Project data for training models is a 3-dimension matrix (M x N x 6)

ecoom
EXPERTISECENTRUM O&O MONITORING

UHASSELT

# Proposed approach

Deep learning models

- Recurrent neural networks

  Long Short-Term Memory Networks (LSTM)

  Bidirectional Long Short-Term Memory Networks (BiLSTM)

ecoom
EXPERTISECENTRUM O&O MONITORING

UHASSELT

# Proposed approach

Deep learning models

- Recurrent neural networks

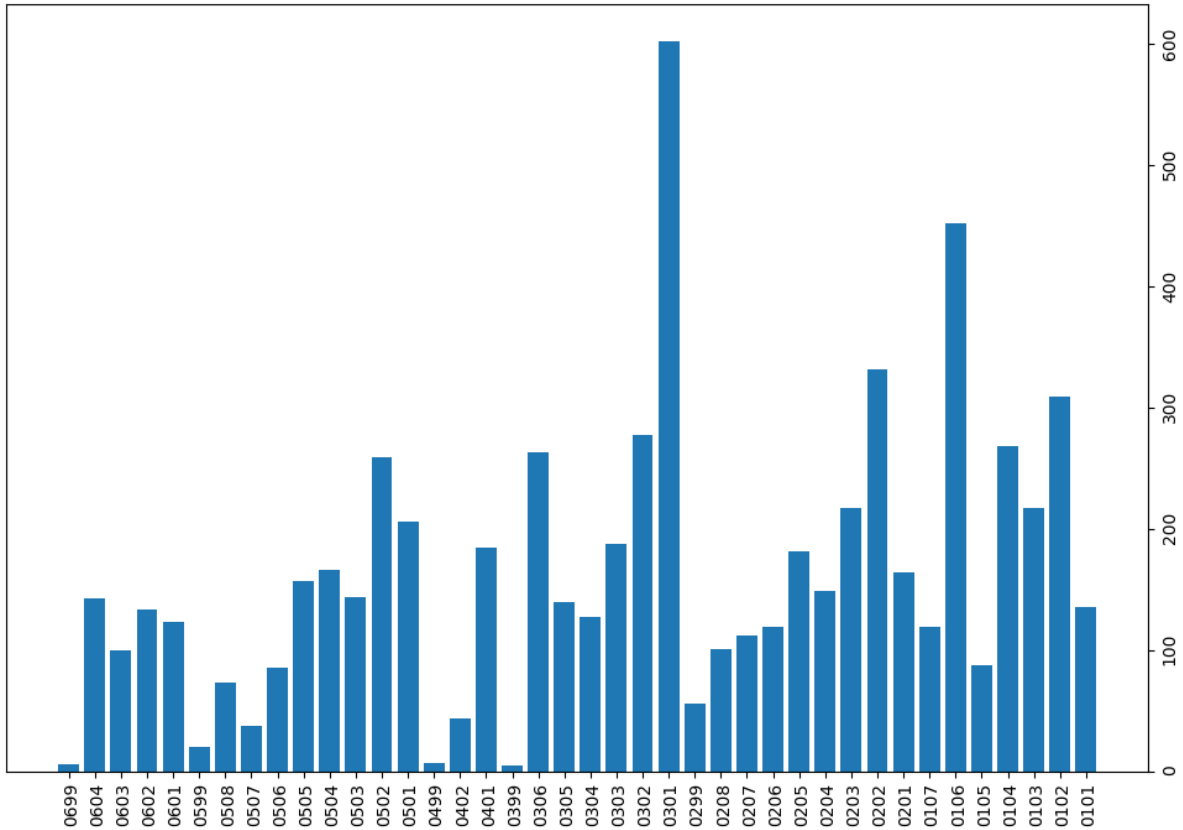
  Long Short-Term Memory Networks (LSTM)

  Bidirectional Long Short-Term Memory Networks (BiLSTM)

- Convolutional Neural Networks (CNN)

# Experimental setup

Dataset: 3954 projects, 42 disciplines
Feature matrix: (3954, 42, 6)
Training and testing: 80%, 20%




ecoom
EXPERTISECENTRUM O&O MONITORING

UHASSELT

# Experimental setup

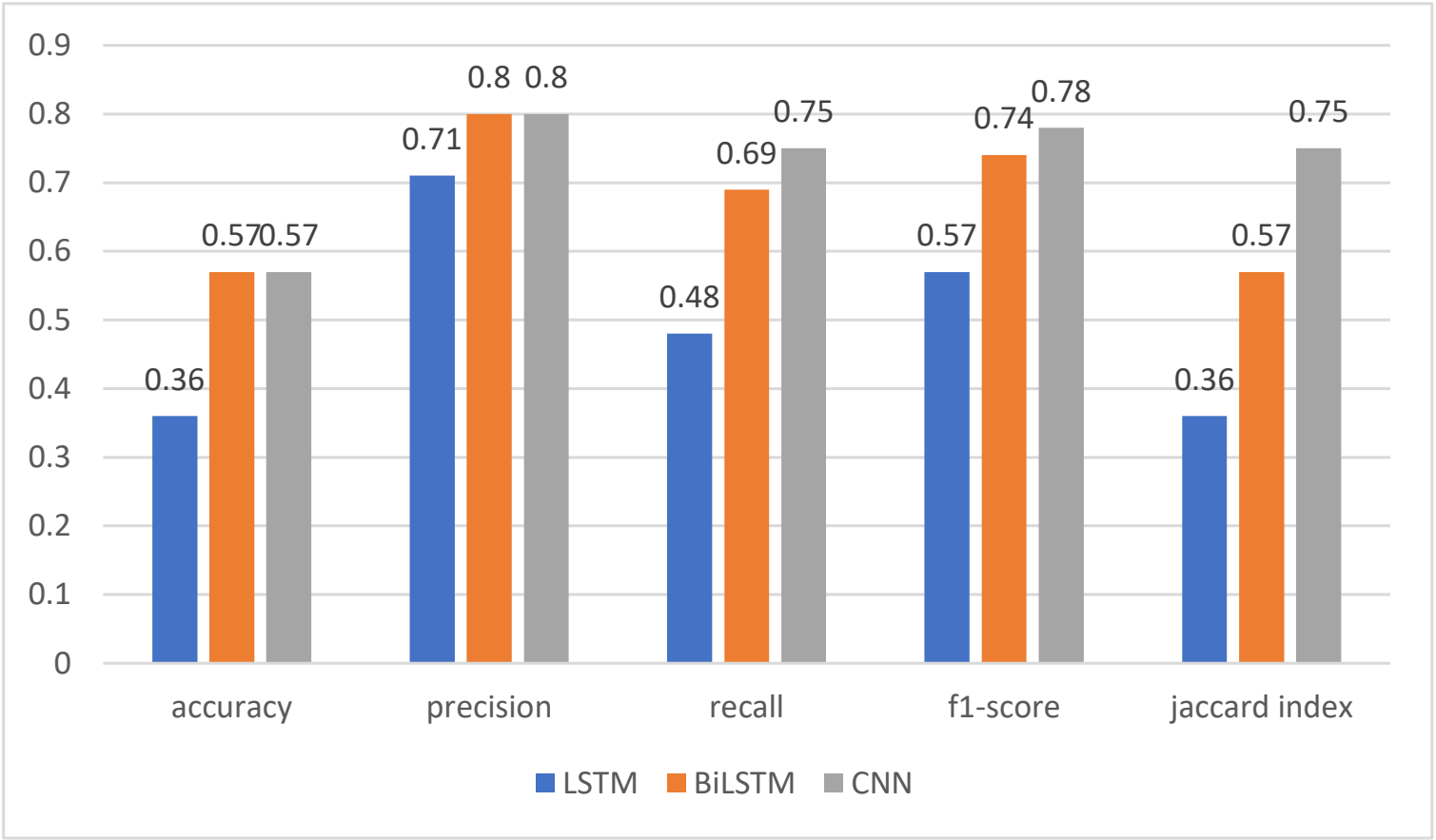
- Model hyperparameters
  - numbers of hidden units = 128
  - epochs = 100
  - optimizer='adam'

# Experimental setup

- Model hyperparameters
  - numbers of hidden units = 128
  - epochs = 100
  - optimizer='adam'
- Metrics
  - accuracy
  - precision
  - recall
  - f1-score
  - jaccard index

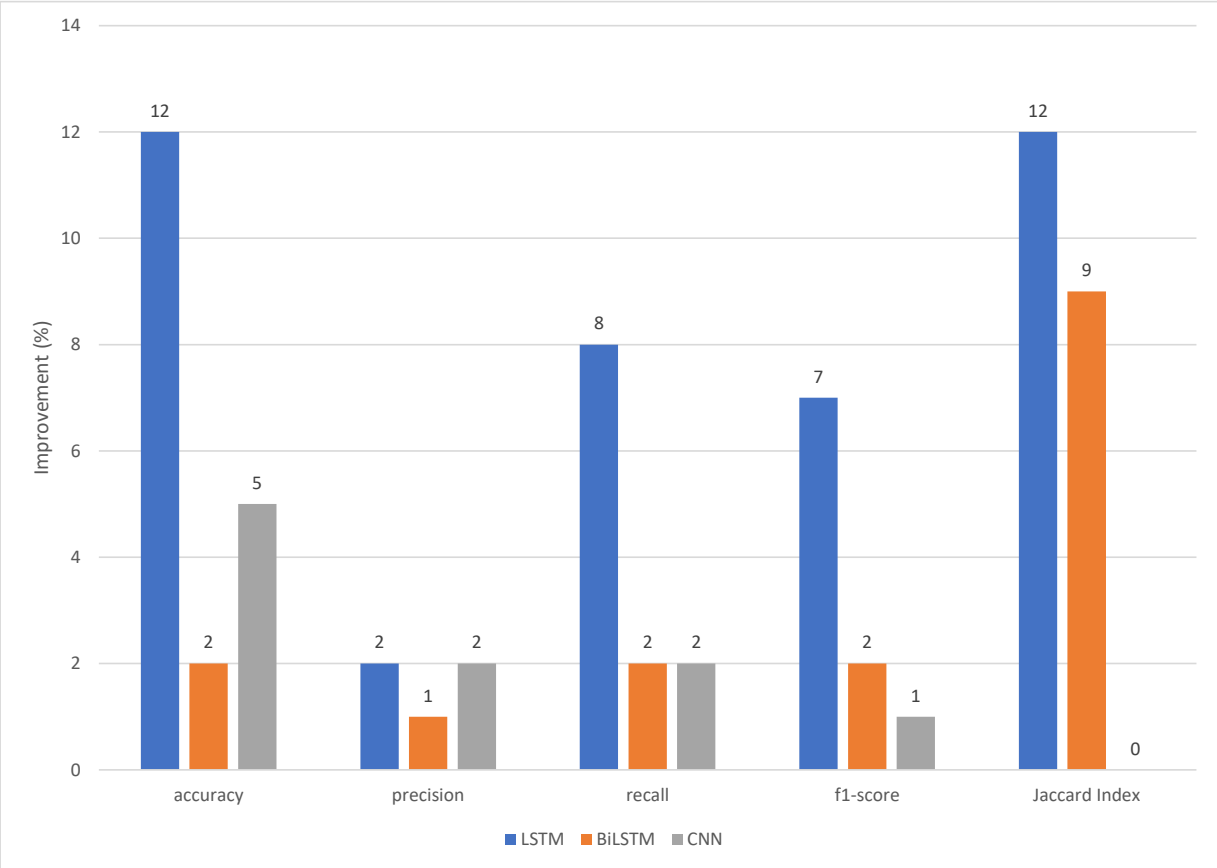
# Performance results

## Performance of models



# Performance results

Performance of the models on the dataset, after excluding projects containing low-frequency disciplines



# Conclusion

We proposed a new approach to predict disciplines related to a project, based on the disciplines of researchers

Performance of deep learning models was relatively high

Further data pre-processing could improve the models' performance

ecoom
EXPERTISECENTRUM O&O MONITORING

UHASSELT

# Conclusion

We proposed a new approach to predict disciplines related to a project, based on the disciplines of researchers

Performance of deep learning models was relatively high

Further data pre-processing could improve the models' performance

Future work

- Further evaluate models on larger datasets
- Experiment with different hyperparameters
- Consider preprocessing the data differently, such as using different normalization techniques or feature engineering.

ecoom
EXPERTISECENTRUM O&O MONITORING

UHASSELT

# Thank you for listening