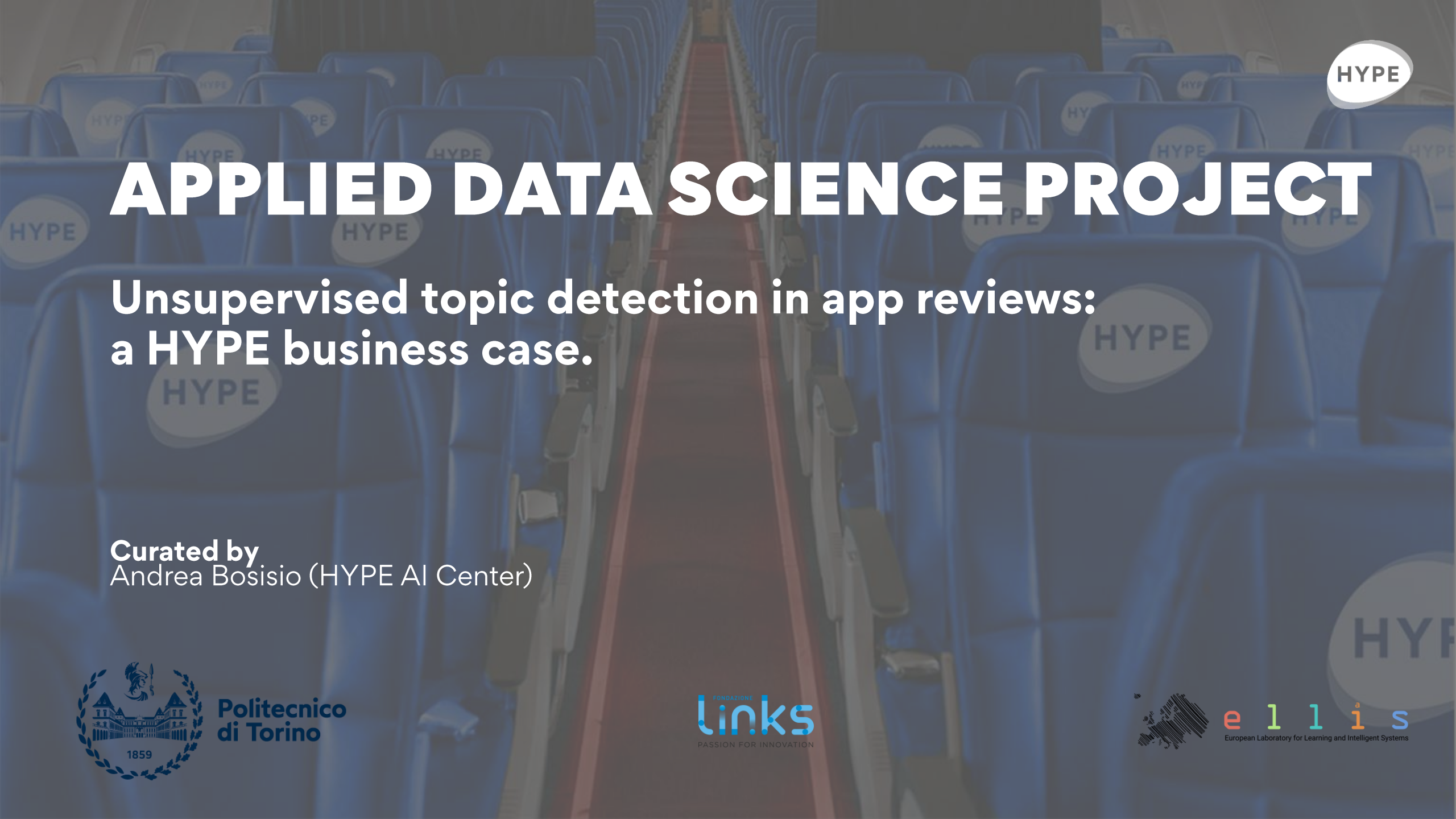


APPLIED DATA SCIENCE PROJECT

Unsupervised topic detection in app reviews: a HYPE business case.

Curated by
Andrea Bosisio (HYPE AI Center)

Politecnico di Torino

Fondazione links
PASSION FOR INNOVATION

ellis
European Laboratory for Learning and Intelligent Systems

A VALUE-DRIVEN PROJECT BORN FROM A REAL HYPE CHALLENGE

How can we **monitor topics in HYPE reviews** to guide future features and improvements?

How to detect topics? How detailed should they be? How to manage evolving or new ones?

How to handle relationships among them?

The Research & Customer Insight team @HYPE is interested in this business case, and the AI Center @HYPE engineered a solution (and is curious to hear new ones 😉)

DATA

The dataset consists of ≈5500 HYPE app's **reviews** (mostly in Italian) collected from various app stores/websites.

Each review includes
- a **rating** (integer: 1...5 or 1...10)
- a **sentiment tag** (either "positive", "negative", or "neutral").

Some of the reviews (≈3800) also come with a **set of labels**.

The **taxonomy** is also provided, containing the description of each label and its possible relationships with other labels.

The dataset will be provided by HYPE as a **CSV/Excel** file.

If needed, other ≈22k raw reviews (no sentiment, no labels) can be provided

HYPE app's reviews and ratings are collected daily from the different sources. HYPE's internal model predicts **sentiment and labels**, which are later validated by the HYPE Research & Customer Insight team.

TASK

The main objective is to develop a method for tagging each review with negative topics (i.e., issues) that should be tracked over time.

The proposed solution should be presented by completing the following tasks:

1. Exploratory Data Analysis (EDA) & pre-processing
2. Topic detection
3. Extra tasks / future works:
   3a. Multi-label classification
   3b. Hierarchical topic detection

1. EDA & PRE-PROCESSING

HYPE

Analyse first the text of the reviews (length, language, ...): are there outliers? Or "junk" reviews?

1

Are all features (rating, sentiment, labels) relevant to the goal?

2

How should positive and neutral reviews be treated compared to negative ones?

3

Perform an initial analysis based on the assigned labels: cardinality, consistency with sentiment and taxonomy, ...

4

Which assumptions can be made to filter out or pre-process samples?

5

2. TOPIC DETECTION

HYPE

Ignore the provided labelling for now:

- Can another reasonable taxonomy (min.10 topics) be generated directly from the data?
- Which topics emerge?
- How well are these topics separated (qualitatively, visually, quantitatively)?
- How can the detection be improved?

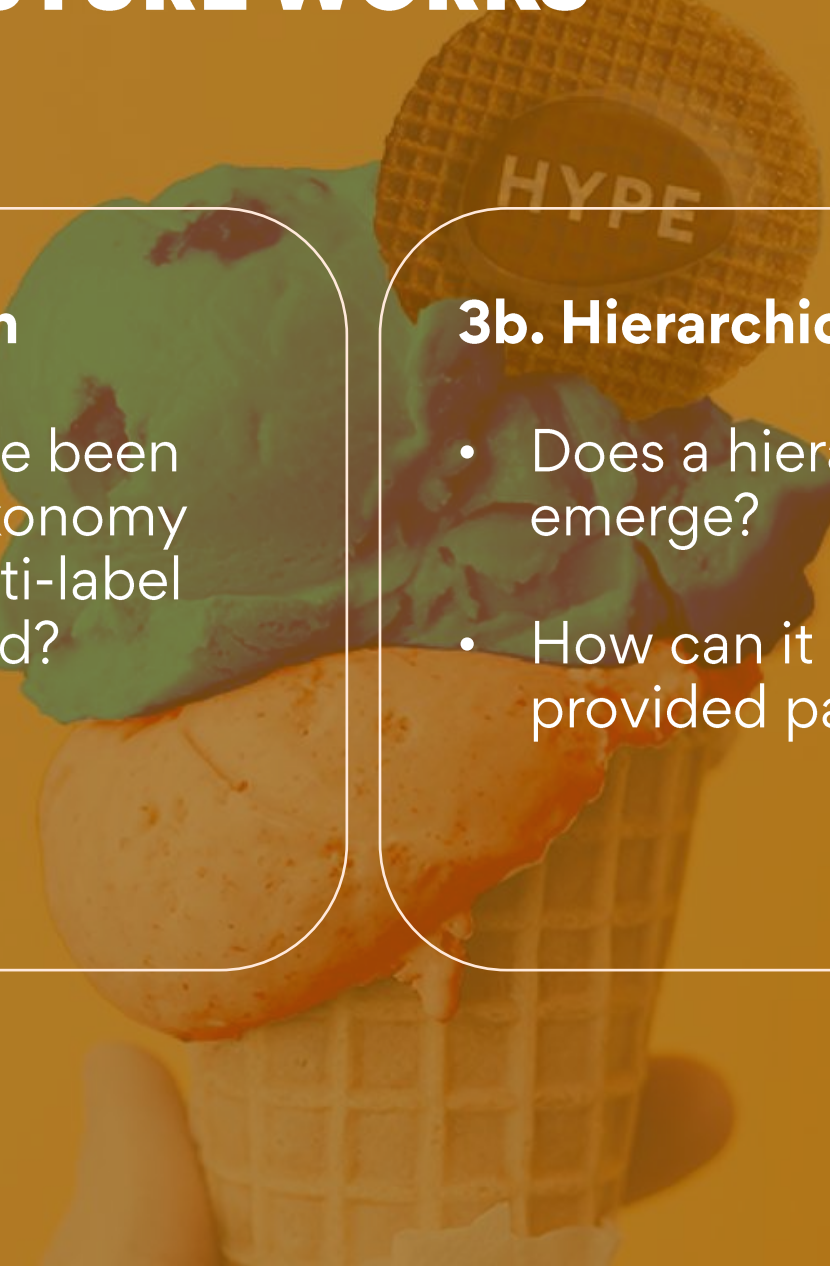
1

Compare discovered topics with the provided taxonomy:

- Is there any overlap? And how can it be measured?
- If overlap is low, how can the detection be guided to align with the provided taxonomy?
- Could the taxonomy be simplified by leveraging the provided parent-child relationships?

2

3. EXTRA TASKS / FUTURE WORKS

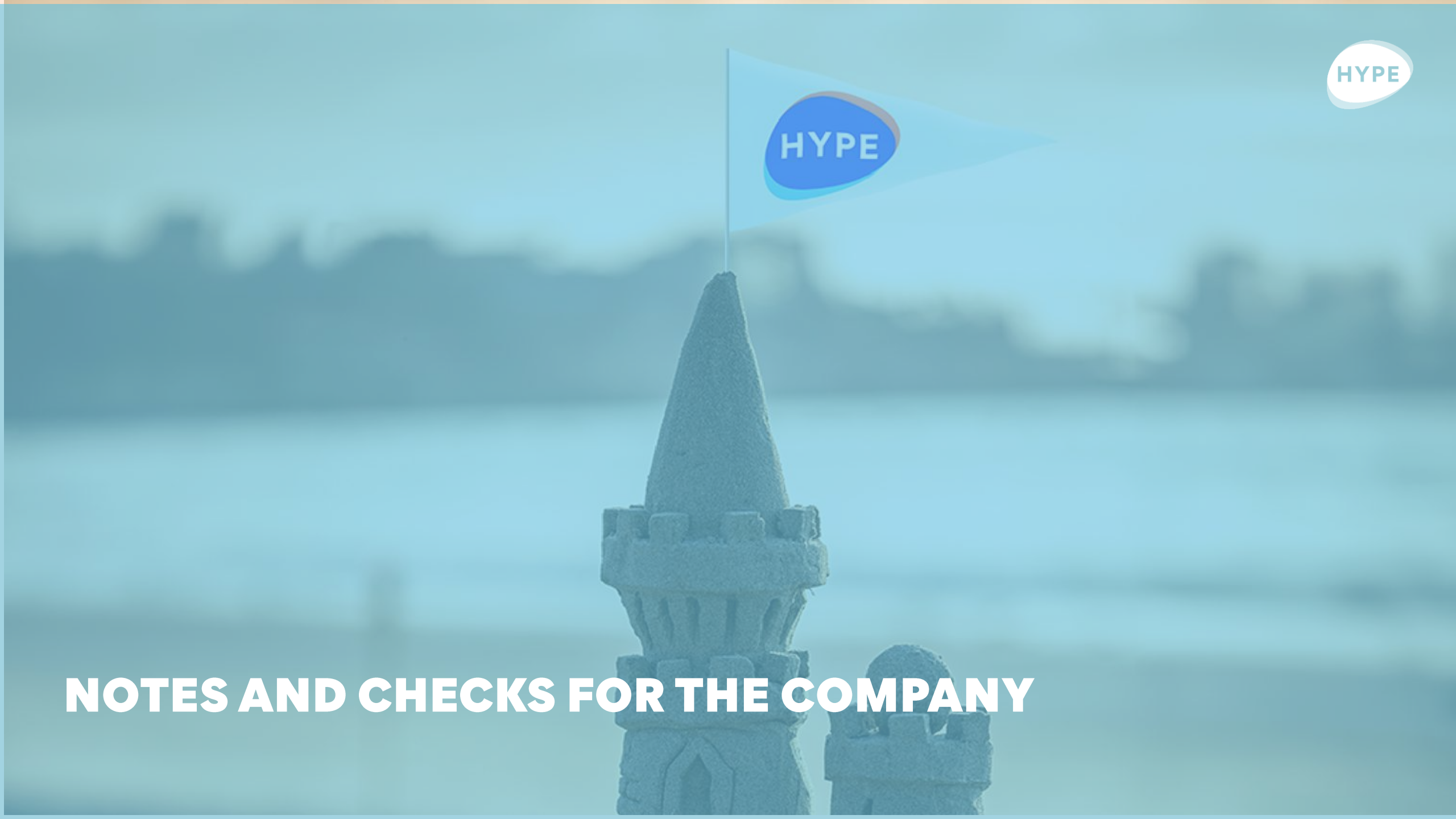


3a. Multi-label classification

- Once detected topics have been mapped to a subset of taxonomy topics (≥10), how can multi-label classification be performed?
- How would you measure performance?

3b. Hierarchical topic detection

- Does a hierarchical structure emerge?
- How can it be compared with the provided parent-child relationships?



NOTES AND CHECKS FOR THE COMPANY

LIGHT MENTORING

Contacts:

- Carla Maria Medoro – carlamaria.medoro@linksfoundation.com
- Andrea Bosisio – andrea.bosisio@hype.it

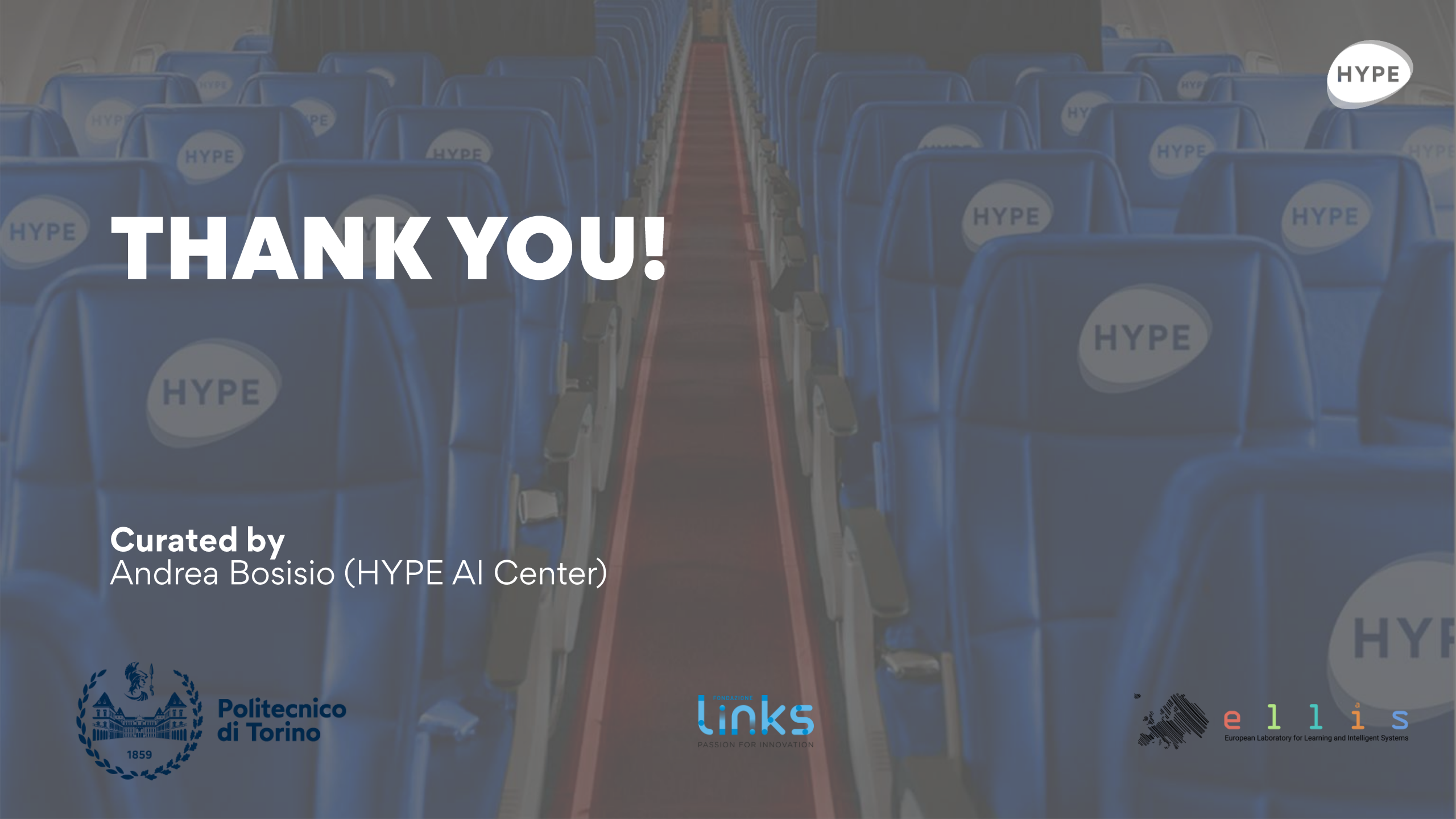
A suggested schema could be having 30 minutes biweekly of calls with students for the whole duration of the semester.

POLICY

Both project descriptions and implementations will be part of a repository group published on GitHub

The company (HYPE) confirms that the project repository can be made public.

Ideally, the projects should be conceived open from the design



THANK YOU!

Curated by
Andrea Bosisio (HYPE AI Center)

Politecnico di Torino

Fondazione links
PASSION FOR INNOVATION

ellis
European Laboratory for Learning and Intelligent Systems