E-MIMIC: Empowering Multilingual Inclusive Communication

Giuseppe Attanasio, Salvatore Greco, Moreno La Quatra,

Luca Cagliero, Rachele Raus, Michela Tonti, Tania Cerquitelli

Department of Control and Computer Engineering, Politecnico di Torino

Department of Interpretation and Translation, Università di Bologna

Data science for equality, inclusion and well-being challenges
Workshop of IEEE BigData 2021 conference

IEEE

ai4ei | Artificial Intelligence for European Integration
Jean Monnet Centre of Excellence

Politecnico di Torino

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DIPARTIMENTO DI
INTERPRETAZIONE E TRADUZIONE

Outline

- Context and motivation
- Challenges
- Proposed pipeline
- Preliminary case study
- Future works

Inclusive Language Problem

- Increasing attention to **inclusive languages**

- Preserving **diversity** and **inclusion**
  - Il parere **dei direttori tecnici** deve essere acquisito entro l'anno
  - Il parere **della direzione tecnica** deve essere acquisito entro l'anno
  - *The opinion of the technical management must be acquired within the year*

- Automated Machine Translation tools increases the non-inclusive text generation problem

Increasing interest in developing automated solutions to preserve inclusivity and diversity in formal communications

Italian Communication Context

- Noticeable problem in romance languages
  - linguistic feminization
  - Stereotypes

- Italian language case study
  - Italian institutions privilege the usage of the masculine as *"neutral"* form

- In 2018, the ministry of education, university and research proposed the set of guidelines for gender-inclusive language
  - Exploiting the guidelines is time-consuming
  - Exists a very large corpus of documents to be rewritten
  - Lot of linguistic expertise is required

→ **Deep learning** could be an effective tool in **spreading inclusive communication**

Deep Learning for NLP

- **Deep Language Models** achieved **impressive performance** on diverse NLP-related tasks

- They acquire knowledge being **trained on large corpora** available on the web, using unsupervised learning

- Each architecture can be applied to solve specific tasks:
  - Encoder models (e.g., text classification)
  - Decoder models (e.g., text generation)
  - Sequence-to-sequence models (e.g., machine translation)

→ Current **Deep Language models** are **not** trained to **model Inclusive Language**

Inclusive Communication Challenges

1. *How* to define **non-inclusive** communication?

2. *Whether* **Deep Learning** techniques are suitable for modeling inclusive language?

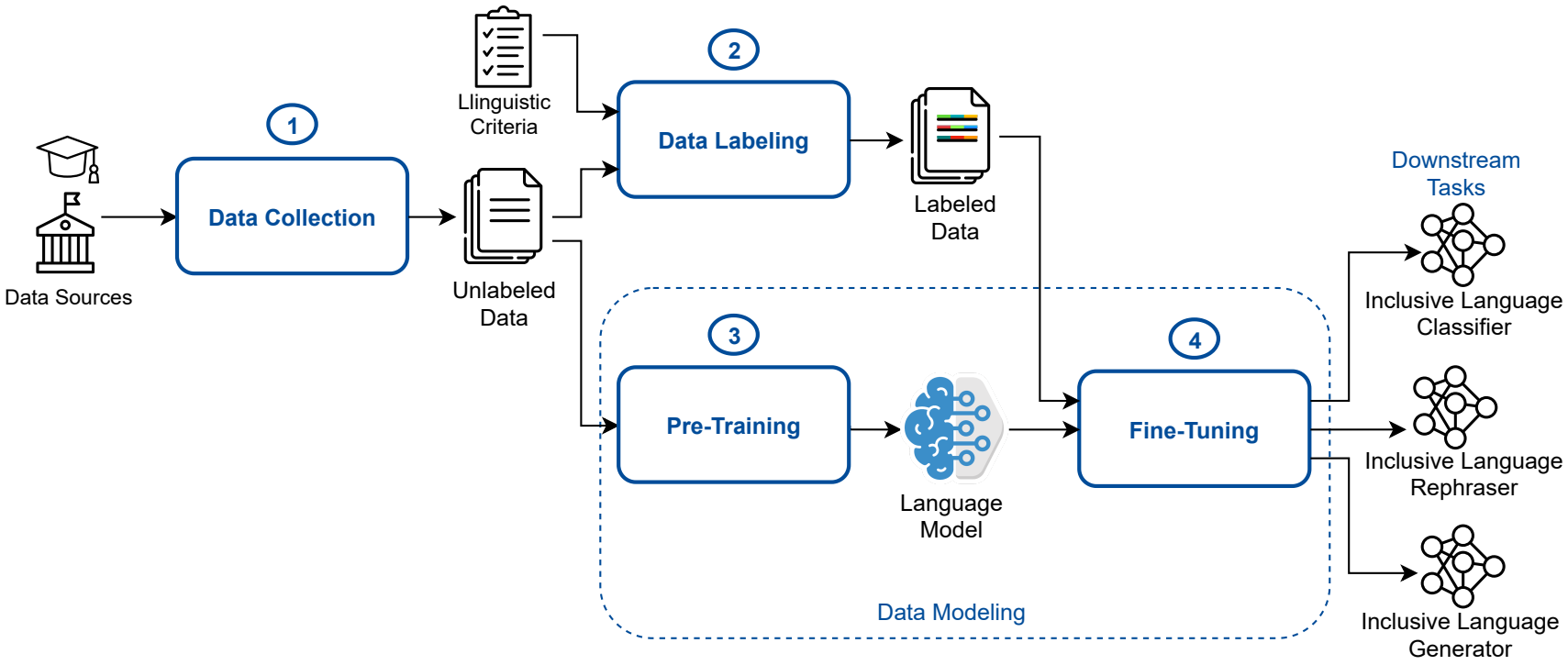
3. *To what extent* **large-scale** collections are suitable for learning pre-trained models?
   - For adaptation to **Inclusive Language tasks**

E-MIMIC

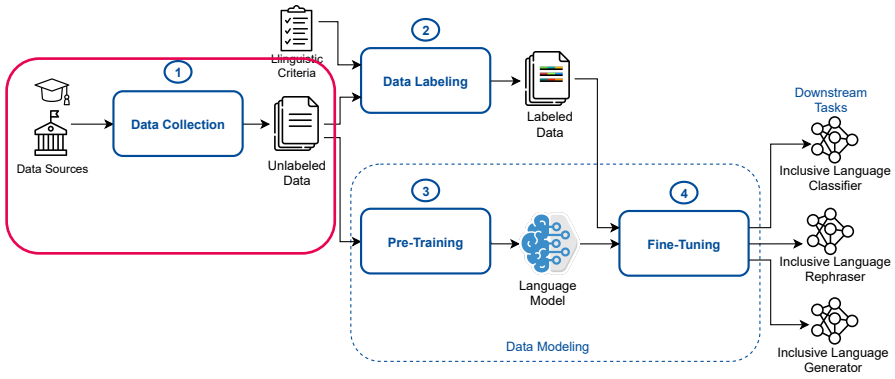
E-MIMIC: **E**mpowering **M**ultilingual **I**nclusive com**M**un**IC**ation

- **Fostering** inclusive communications in real-world scenarios
- **Detecting** and **overcoming** language inclusivity issues
  - Grammatical asymmetry (silencing the feminine form)
  - Semantic asymmetry (presence of stereotypes)
- Exploiting a **deep learning** pipeline to **generalize** and **automate** the process
  - It leverages language models trained on purpose-specific corpora
  - It can detect non-inclusive expressions and suggest inclusive alternatives
- Currently focusing on **Academic** and **Public Administration Italian** documents
  - Can be adapted to different languages and types of documents

E-MIMIC Pipeline



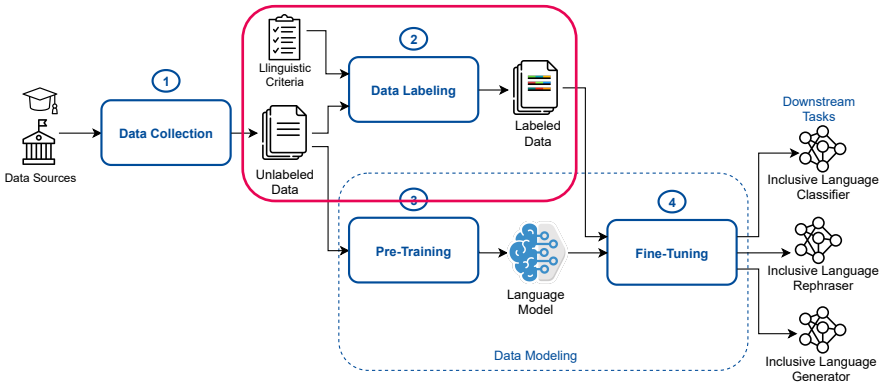
1) Data Collection



- **Limited** amount of **annotated data** for romance languages
  - **Italian** is particularly prone to **non-inclusive language** (first case study)
  - No publicly available dataset annotated for inclusive Italian language detection
- Focus on Inclusive language in:
  - Administrative documents
  - Grants
  - Internal and external policies
  - Calls for applications
- Easily **extendable** to different types of documents

→ Accurate **data collection** and **annotation** is of primary relevance for high-quality results

1) Data Labeling



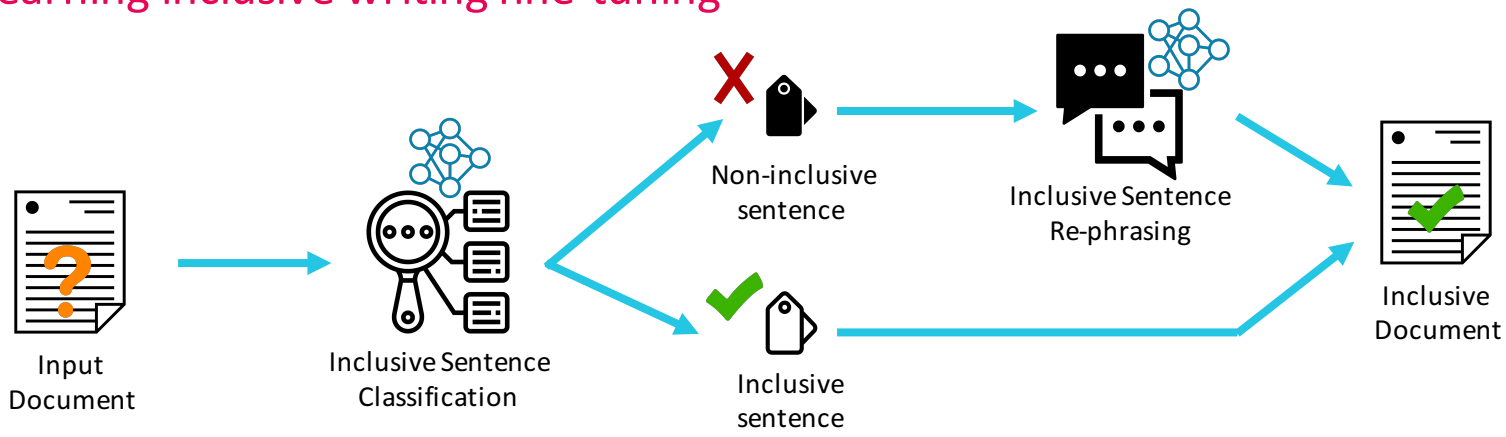
1. Sentence-level splitting
   - Shortest unit containing non-inclusive phrasing
2. Sentence annotation
   - Inclusive label
     - Inclusive, Non-Inclusive, Non-Applicable
   - Part-of-speech tagging of salient linguistic features
     - Cited content, proper names, potentially stereotypical phrases, etc.
   - Type of content
     - Legal, administrative, technical, or informative document
   - Edited sentence part highlighting
   - Inclusive re-phrasing formulation of the sentence
     - One or more valid inclusive re-formulation

→ **Linguistic criteria** design and **annotation** is performed by **Linguistic Experts**

3) Data Modeling

1. Target domain/task specialization
   a) Masked Language Model (MLM) for administrative and academic language specialization
   b) Named Entity Recognition (NER) on Part-of-speech for inclusivity tags
   c) Content classification (i.e., predict whether a sentence contains legal, administrative, technical, or informative content)

2. Learning inclusive writing fine-tuning



Preliminary Case Study

Synthetic dataset

- Inclusive sentence classification (Inclusive *I*, non-inclusive *NI*)

- Template filling procedure
  - Collected 19 templates and 43 seeds
  - producing 822 annotated Inclusive *I* and non-inclusive *NI* samples

Template	**IT:** Occorre richiedere la firma **[blank]** **EN:** One must request the signature of **[blank]**
Syntetic examples	***NI*:** Occorre richiedere la firma degli interessati ***I* :** Occorre richiedere la firma delle persone interessate **EN:** One must request the signature of interested people

Preliminary Case Study

Sentence inclusivity classification

- Randomly selected 80% of the templates for training
  - 70% as training set
  - 10% as validation set
- Remaining 20% of templates as test set
- Templates restricted to either training or test set
  - prevent the model from identifying shallow syntactic patterns
- Fine-tuned a pre-trained BERT checkpoint for Italian language
  - 85% of accuracy on test set (i.e., new unseen templates) for sentence inclusivity classification

→ Promising ability of **deep language models** to effectively **represent inclusive language**

Future Works

- **Inclusive data annotations**
  - Existing documents are manually annotated to detect non-inclusive language
  - Manual rephrasing is proposed for non-inclusive sentences

- **Inclusive language modeling**
  - Deep learning-based language models are trained with a mix of self-supervised and supervised approaches
  - They are trained to solve multiple tasks for the goal of inclusive communication (e.g., classification and reformulation)

- **Multilingual pipeline adaptation**
  - The pipeline proposed for Italian language can be adapted to multiple languages (e.g., other romance languages as French or Spanish)

Thank You!

Data science for equality, inclusion and well-being challenges
Workshop of IEEE BigData 2021 conference