CC BY NC SA

Software projects

Karën Fort and Julien Bezançon

karen.fort@loria.fr / https://members.loria.fr/KFort

Institut des sciences du Digital Management & Cognition
COMPOSANTE DE L'UNIVERSITÉ DE LORRAINE

UNIVERSITÉ DE LORRAINE

Objectives

- Put into practice NLP tools and methodologies adapted to the question at hand in a real-world setting:
  - datasets not always available
  - computing power not always available
  - the language is not always English
- Creating data, when necessary
- Applying sound development practices: documentation, testing, ethical analysis
- Evaluating systems

Course organization

10 sessions, incl. this one and the final one with your presentations

Use:

- GitLab repository (use the one from the university, https://gitlab.univ-lorraine.fr/)
- Overleaf

Course evaluation

What will be taken into account:

- the right tool for the task: do not use LLMs unless absolutely necessary (and prove it)
- avoid English (unless the topic is really original)
- documentation, clarity of the code
- evaluation of the tool and its limits

- grade: final presentation + repository and overleaf (doc)

Organization

- 3 students (a couple of groups of 4): I need an **even number of groups**
- subject:
  - your own choice (should be made by the end of the second course) OR
  - from our propositions OR
  - a variation from our propositions (language, corpus, etc)

Analysis of climate denial discourse from politicians (French)

Goal: automatically identify climate denial discourse

- in French politicians transcribed speech from Assemblée nationale (corpus provided as csv)
- using the classification provided by https://www.nature.com/articles/s41598-021-01714-4

Tasks/requirements:

- state of the art (who worked on that? what did they say?)
- gather statistics about the corpus (and check the corpus for potential errors)
- perform an automatic extraction of climate denial discourse and identify by whom

Evaluation of ableist biases of LLMs (languages != English)

Goal:

- develop a tool/method to evaluate stereotypical biases wrt handicap in LLMs

Tasks/requirements:

- State of the art
- Find an idea ;-) : you can take inspiration from resume screening experiments[1], or you can generate cover letters, for ex
- Experiment building a tool to automatically detect ableism
- Evaluate the results wrt state of the art

[1] e.g., [Glazko et al, 2024]: https://dl.acm.org/doi/pdf/10.1145/3630106.3658933

Automatic detection of transphobia

Goal: develop a tool to automatically detect transphobic comments (for example on YouTube videos)

- open
- French (or language TBD, not English or Tamil)

Tasks/requirements:

- Building a corpus
- Manual analysis/annotation of transphobic content (can be offensive)
- Experiment building a tool to automatically detect transphobia
- Evaluate the results wrt state of the art

Evaluation/detection of socio-economic stereotypical biases in LLMs

Goal: develop a tool/method to evaluate socio-economic (classist) stereotypical biases in LLMs

- very open
- any language

Tasks/requirements:

- State of the art
- Find an idea ;-) : you can take inspiration from *The AI GAP* paper[2]
- Experiment building a tool to automatically detect classism/aporophobia
- Evaluate the results wrt state of the art

[2] [Bassignana et al, 2025] https://aclanthology.org/2025.acl-long.914/

She's a woman, **but** *she can drive*: linguistic linkers as stereotypes clues

Goal: Use linguistic linkers (*but, however, yet, so, ...*) to detect implicit stereotypical bias (of any type!)

- open
- any language (as long as it has equivalents of linguistic linkers)
- can be conducted on LLM-generated content, and/or on any other corpus
- potential internship

Steps for the project:

- Choose/create/generate a relevant corpus
- Detect/collect linguistic linkers
- Detect and measure the presence of stereotypes
- Tackle semantic phenomena: contradictions, mutually exclusive terms, paradoxes, ...

Analysis (?) of the Grand Débat corpus (French)

Goal: whatever analysis you want to perform on the Grand Débat corpus (https://granddebat.fr/pages/donnees-ouverteshttps://journals.openedition.org/revuehn/2655)

- open
- French

Steps for the project:

- get to know the corpus
- state of the art of what's been done on it
- what's missing/interesting to evaluate
- develop tools for this

Analysis (?) of Spontaneous Dialogues in L1 English corpus

Goal: whatever analysis you want to perform on the corpus (https://www.ortolang.fr/market/corpora/cles-en_corpus)

- open
- English

Steps for the project:

- get to know the corpus
- state of the art of what's been done on it
- what's missing/interesting to evaluate
- develop tools for this

Analysis of a multilingual news articles corpus

Goal: whatever analysis you want to perform on a corpus consisting of millions of headlines and summaries of news articles (https://www.kaggle.com/datasets/felixludos/babel-briefings).

- Multilingual
- study at least two languages

Steps for the project:

- get to know the corpus
- state of the art of what's been done on it
- what's missing/interesting to evaluate
- develop tools for this

Automatic identification of snowclones in scientific article titles

Goal: Identify snowclones (wordplays derived from an expression or a movie title) in a dataset of scientific article titles (extracted from OpenAlex https://openalex.org/)

- dataset to collect (API requests)
- any language (easiest = English)

Steps for the project:

- collect and explore the corpus
- create a list of expressions or movie titles (e.g.: the good, the bad and the ugly)
- find snowclones of said expressions in the corpus (e.g.: the good, the bad, and the **tuna**)
- develop tools for each step

Automatic snowclone generator

Goal: create a snowclone generator using the FROST lexicon (bezançon et al., 2026), a snowclone lexicon containing movie titles.

- dataset to collect (reproduce bezançon et al. experiment https://github.com/JulienBez/LSHSnowclone)
- Multilingual

Steps for the project:

- collect and explore the corpus
- isolate snowclone patterns (e.g.: for a fistful of X, X being an open slot for any noun)
- Isolate a list of terms to use in open slots (e.g.: for a fistful of **dollars**, for a fistful of **kebabs**, for a fistful of **seeds**)
- generate new snowclones in multiple languages
- develop tools for each step

LLMs for health: uncovering the hype in research papers

NLP for NLP - argumentative/rhetorical analysis - discourse analysis

Context: LLMs are increasingly taking up space in NLP conferences and publications (80% of ACL 2025 papers!); the text of these publications can be used to analyze the representations of LLMs that researchers convey to the community.

Objective: To study the discourse on LLMs within research papers focused on health applications: which topics, which promises, which framing of LLMs?

Tasks: You will have to parse papers (ACL Anthology, PubMed - provided!), and:

- Apply methods for claim category identification and anthropomorphization detection (guided)
- Propose methods to identify topics and recurring promises (possibly along manual annotation) → or any other relevant idea you have!

Subject proposed by Clémentine Bleuze (*clementine.bleuze@univ-lorraine.fr*)

LLMs for health: extracting smoking status from clinical documents (FR)

Context: textual narratives are prevalent in electronic health records (80% of all EHR data!); information extracted from this text can be used to assist clinicians in retrospective studies.

Objective: analyze documents from the PARHAF corpus and extract smoking status

Sample text: Tabac : 5 c environ depuis janvier (mais environ 30 PA en tout)
Analysis: Tabagisme actif; 30 PA

Sample text: a fumé 5 cig par jour pendant 10ans
Analysis: Tabagisme sevré; 2,5 PA

Sample text: Pas de tabagisme.
Analysis: Tabagisme négatif; -

Subject proposed by Aurélie Névéol (*aurelie.neveol@lisn.fr*)

Measuring the presence of BigTech in the Speech community

Goal: adapt the work done on NLP [Abdalla et al., 2023] to the speech processing community (Interspeech conference)

- existing archive of the Interspeech papers (unique)
- collaboration with Sébastien Le Maguer

Steps for the project:

- check the corpus
- do an automatic analysis of the affiliations of the authors
- perform whatever complementary analysis (topics of interest for BigTech, etc)

Subject proposed by Vincent P. Martin and I (publication planned)

Documenting who is doing which project

First arrived, first served! (do not take a project that is already taken)

https://lite.framacalc.org/j158c0uq3t-anad

Abdalla, M., Wahle, J. P., Ruas, T., Névéol, A., Ducel, F., Mohammad, S. M., and Fort, K. (2023).
The Elephant in the Room: Analyzing the Presence of Big Tech in Natural Language Processing Research.
In 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada.