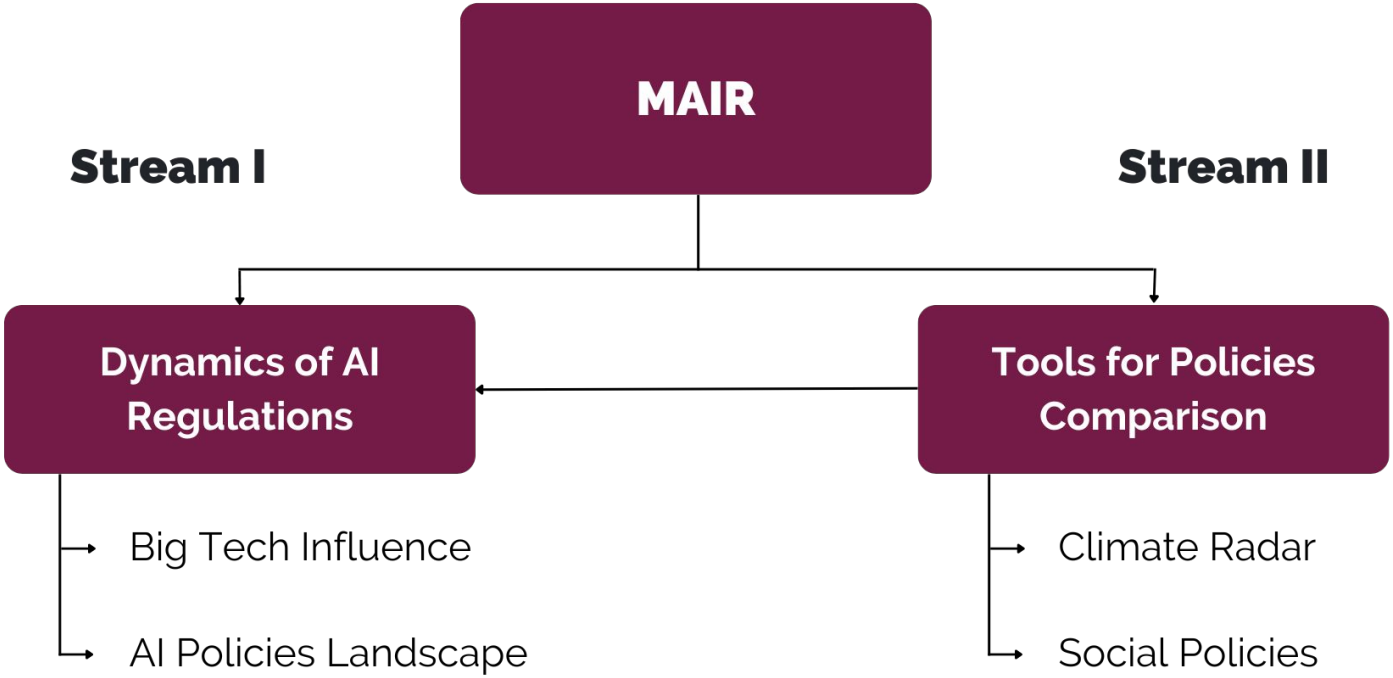
ATLAS

Automated documen**T** ana**L**ysis for soci**A**l awarene**S**s

MI²DataLab Winter Seminar 2022

MI

Create qualitative and quantitative **NLP tools** for efficient and **automated analysis of documents** to **increase social responsibility** and awareness


MAIR
Stream I
Stream II
Dynamics of AI Regulations
Tools for Policies Comparison
Big Tech Influence
AI Policies Landscape
Climate Radar
Social Policies

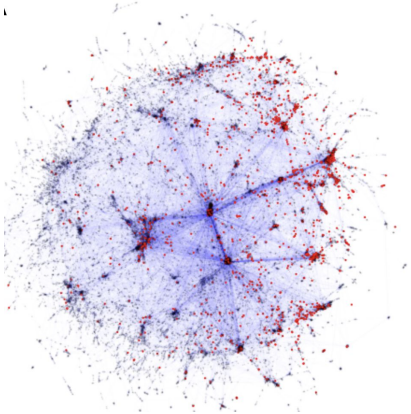
Stream I

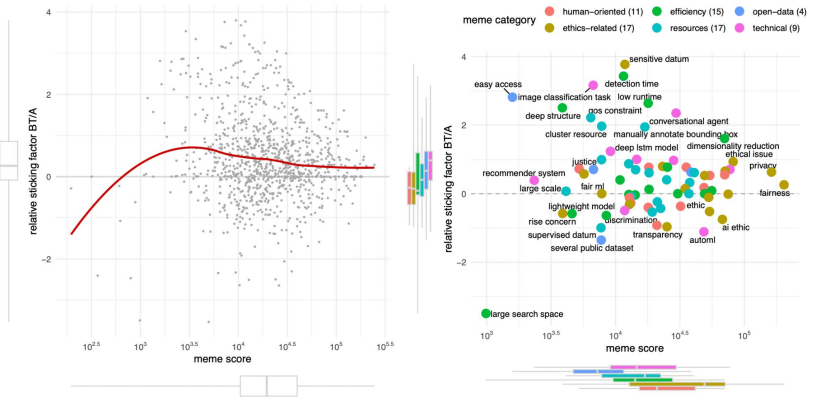
Uncovering Influences and Dynamics in AI Regulations

Recap: How does Big Tech influence AI research?

In this work, we wanted to understand which ideas are spread by big tech companies in AI research papers.

We leverage NLP to extract ideas from papers in the network, and then measure their “infectiousness” depending on who is talking about them.



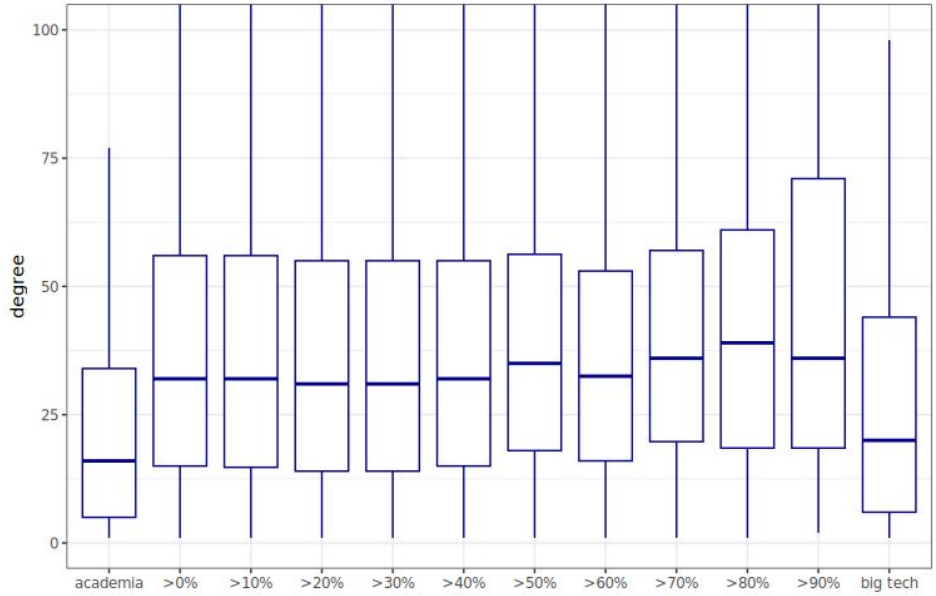


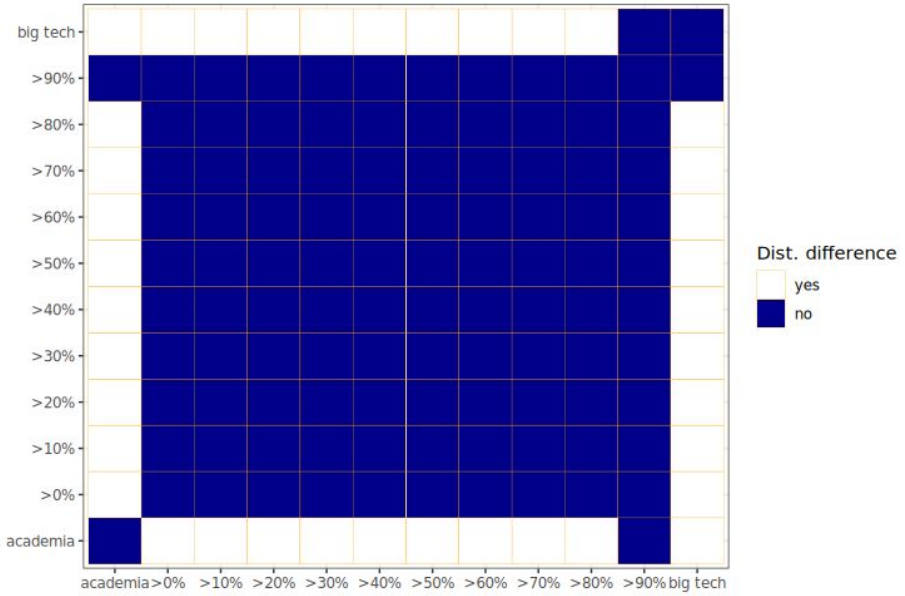
Stanisław Giziński

Challenges and findings

1. Big tech and academia papers not-binary distinction
2. Big tech vs other companies – is the distinction needed
3. Manual validation of memes

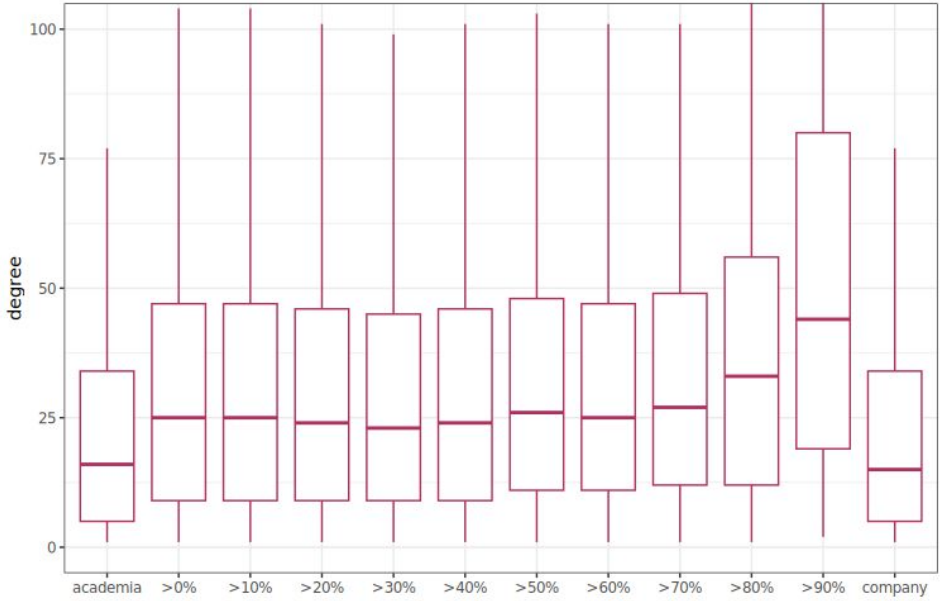
Big tech vs academia network analysis

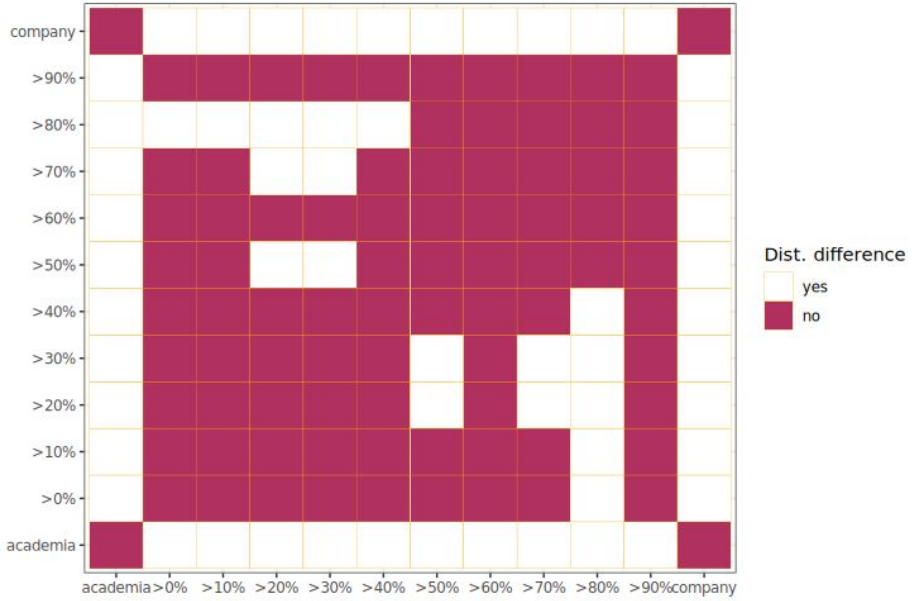




Stanisław Giziński

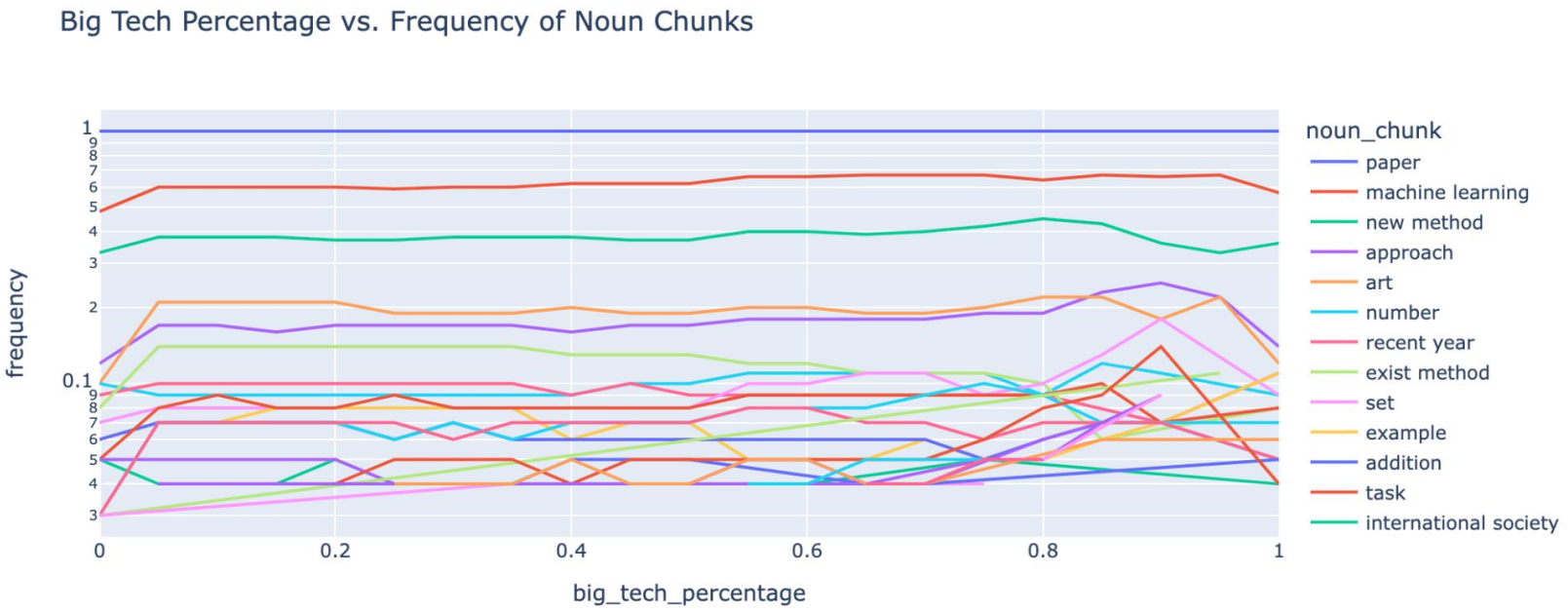
Company vs academia network analysis





Stanisław Giziński

Chunks frequency – how it changes with big tech/academia threshold



Stanisław Giziński

Validation strategy

Our method of finding memes adds 2 layers on top of original meme score method.

1. Noun chunks extraction
2. Embedding+clustering

We want to validate both layers separately, to check how each layer improve memes quality.

As a validation metric, we choose:
1. Number of feasible memes in top 100 memes
2. Number of duplicates in top 100 memes

Manual validation

meme_score	meme_name_common	Hubert			
		Common sense	AI technology	AI use case	Score
13076403092	weighted selection	1	1	0	2
11623218841	yarn color	0	0	0	0
11623218841	project sample	0	0	0	0
11623218841	vary illuminant	0	0	0	0
11623218841	component wise and error cascade perspective	1	1	0	2
11623218841	here phase single stage topology	0	0	0	0
11623218841	various visual theory	0	0	1	1
11623218841	strong external signal	0	1	0	1
11623218841	give new task	0	1	0	1
11623218841	input information cognitive and emotional parallel streaming method	1	1	1	3
11623218841	information asset	1	1	0	2
11623218841	sparse and inconsistent code mix datum	1	1	0	2
11623218841	computationally few expensive regard execution time	0	1	0	1
11623218841	fixation assist module	1	1	0	2
11623218841	public health expert	1	0	1	2
11623218841	neural network architecture exploration	1	1	0	2
11623218841	rcl algorithm	1	1	0	2
11623218841	the storage profile	1	1	0	2
11623218841	datum analysis study	1	1	0	2
11623218841	relative complexity	1	0	0	1

Stanisław Giziński

Stream II

Tools for Efficient and Automated Analysis of Documents

MI

Policy Comparison - quick recap

Project genesis: Case Study course, track: NLP in social sciences

Motivation:

1. Increasing number of policy documents
2. Tediousness of manual analysis of documents
3. Limited citizen governance and restricted accessibility for the society

Objective:

Creating **a set of tools** (the entire **pipeline**) that will allow to c**reate a comparative analysis of documents** with a strictly defined structure, utilising clearly defined format.

Emilia Wiśnios

What changed?

OpenAI Climate Change Hackathon

What we've done:

1. **Frontend** changes in the application
2. Adding **summarization** module
3. Prototype of **contextualized** topic modeling model

We've also published a **dataset** of NECPs documents on Kaggle.

EMILIA WIŚNIOS AND 1 COLLABORATOR · UPDATED 2 MONTHS AGO

27 | New Notebook | Download (5 MB)

National Energy and Climate Plans (EU)

Textual dataset extracted from NECPs and divided to sections

DATASET STATS

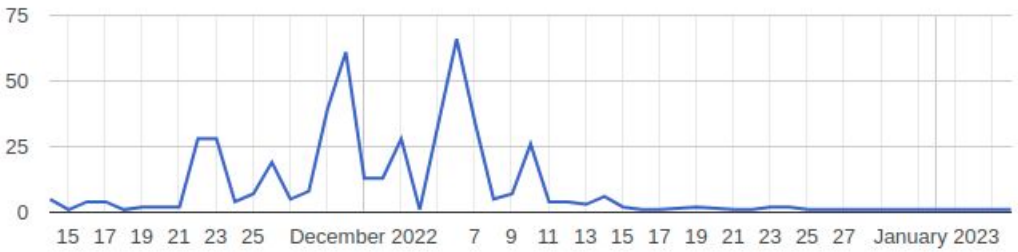
VIEWS
3889

DOWNLOADS
445

DOWNLOAD PER VIEW RATIO
0.11

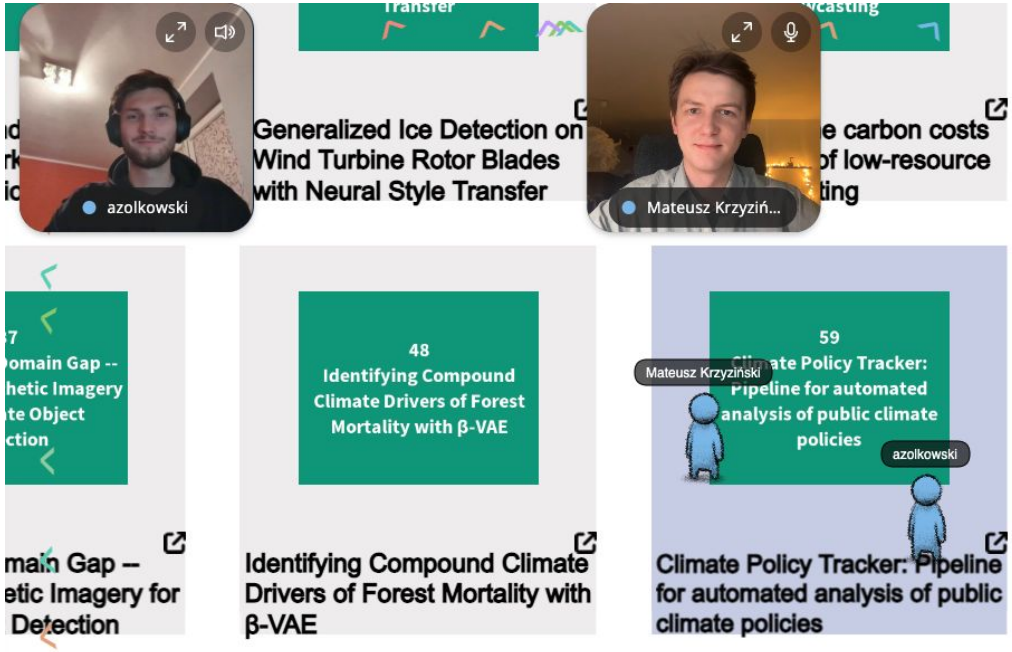
TOTAL UNIQUE CONTRIBUTORS
0

Downloads



Emilia Wiśnios

NeurIPS Workshop on Tackling Climate Change with Machine Learning



Emilia Wiśnios

Data Science Summit 2023

Emilia Wiśnios

Policy Comparison: further plans

- Paper for ACL Demo track (deadline in February)
- Potential cooperation with a client (in progress)

Detecting tensions in UNESCO Proceedings

Project genesis: Bachelor thesis, Computer Science major at MIMUW

Motivation:

1. No existing studies on textual analysis of UNESCO proceedings.
2. New methods in argument mining field.
3. Huge, unexplored dataset for studying argument mining.

Joanna Wojciechowska

Detecting tensions in UNESCO Proceedings

Update:

1. Webscrapping — fetching all the data from whc.unesco.com/en/sessions and whc.unesco.com/decisions. Create metadata.
2. OCR of the documents.
3. Data clearing, split into paragraphs.
4. Data exploration:
    - Most popular n-grams.
    - Topic modelling.
5. First model for controversy detection.

Joanna Wojciechowska


world heritage
state party
heritage committee
heritage centre
draft decision
th session
outstanding universal
states parties
advisory bodies
heritage list
state conservation
universal value
chairperson thank
list world
heritage danger
national park
site
buffer zone
unesco
inscribed world
following
considered
session committee
nomination
mr chairperson
requests state
development
issue
area
important
secretariat
heritage fund
recommendation
stated
including
natural heritage
monitoring mission
programme
particular
summary record
decision comadopted
cultural natural
convention
case
new
region
discussion
view
matter
term
work
time
united state
general assembly
proposed
situation
change
heritage site
working group
place
management plan
iucn
united kingdom
year
point
heritage properties
session world
way
committee member
russian federation
south africa
cultural landscape
activities
thank madam
concerning
regard
question
observer delegation
heritage convention
process
meeting
conservation report
cultural heritage
result
support
need
operational guideline
patrimoine mondial
delegation united
future
project
madam chair
periodic reporting
proposal
thank mr
reactive monitoring
thank chairperson
include
conservation property
committee th

Detecting tensions in UNESCO Proceedings

Future work:

1. Manually checking the results of the model.
2. How to improve it?
3. Which other models can we try out?
4. Repeat.

Joanna Wojciechowska

Inspired Theses

Extending Our Knowledge Within and Outside MI²DataLab

MI

Explainable abstractive summarization of legal acts

Author: Emilia Wiśnios

Collaboration with **Inez Okulska, PhD** from NASK

Motivation:

Amount, structure and language of legal acts are **difficult to understand for people**. We want to make tools for responsible summaries of those documents.

Explainable abstractive summarization of legal acts

What is done?

- Code for extractive summarization using coreference (testing benchmarks in progress, paper writing soon)
- Attacks on GPT3 (experiments in progress)

Emilia Wiśnios

Questions?

MI