CASE 1: HOPEING TO FAIR WELL

Kristoffer Nielbo kln@cas.au.dk

AARHUS
UNIVERSITY
CENTER FOR HUMANITIES COMPUTING

CENTER FOR
HUMANITIES
COMPUTING

OUTLINE

1. During COVID-19

HOPE
FAIR ambition
Media monitoring

2. Post COVID-19

The unstructured challenge
Models cards and data sheets

"Open Science is a complex matter, where the demands for openness and transparency are difficult to comply with."

OUTLINE

1. During COVID-19

- HOPE
- FAIR ambition
- Media monitoring

2. Post COVID-19

- The unstructured challenge
- Models cards and data sheets

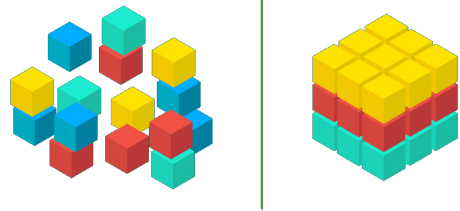
How Democracies Cope with COVID19

How do democracies react and cope as the COVID-19 crisis unfolds and with what effects?

- trajectory of the COVID-19 pandemic
- decisions of governments and international organisations
- decisions of media and social media landscapes
- citizens' behavior and well-being

Open Science and FAIR ambitions from day one

Dana Schutz's 'Building the Boat While Sailing', source: The New Yorker, 2012

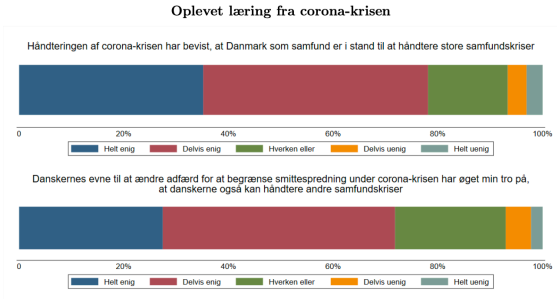


Types of data, source: Coveo, 2022

Public facing information

Three modes of data-driven public communication

- News-like website: `hope-project.dk`
- Interactive dashboard: `hope-project.dk/dashboard/`
- HOPE reports: `hope-project.dk/reports`



All data were stored and shared internally in a GDPR-compliant VRE and a shared data processing agreement

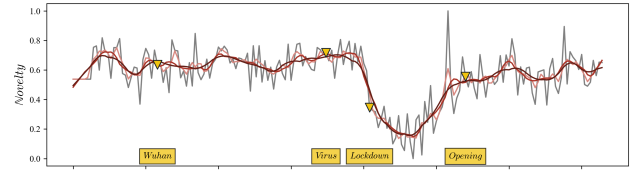
Open (structured) data

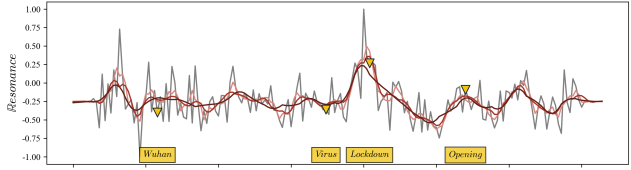
- Data pushed in regular intervals and available upon request
- All survey materials and data deposited in a publicly accessible database (OSF and Zenodo)
- FAIR ambition using a minimal DCAT v2 template

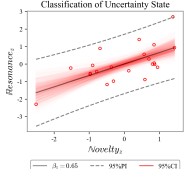
Open source

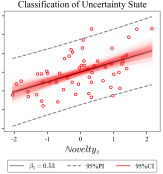
Proto FAIR software approach:

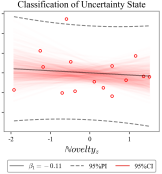
- GitHub - publicly accessible repo with version control
- MIT license - added license
- Zenodo - registered in community registry
- `CITATION.cff` - citation of software
- online sustainability evaluation - quality checklist

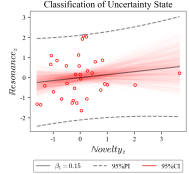

Novelty
Wuhan
Virus
Lockdown
Opening
12-02
12-27
01-21
02-15
03-11
04-05
04-30
05-25

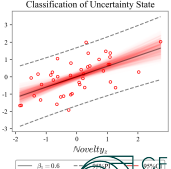

Resonance
Wuhan
Virus
Lockdown
Opening
12-02
12-27
01-21
02-15
03-11
04-05
04-30
05-25

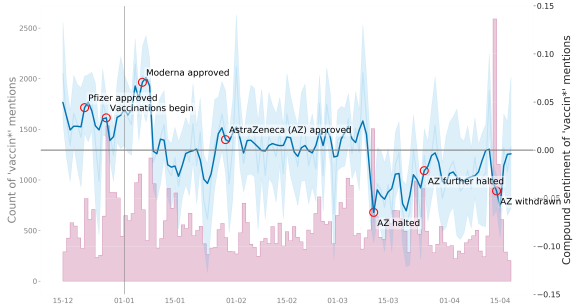

Classification of Uncertainty State
Resonance$_t$
Novelty$_t$
$\beta_1 = 0.65$
95%PI
95%CI

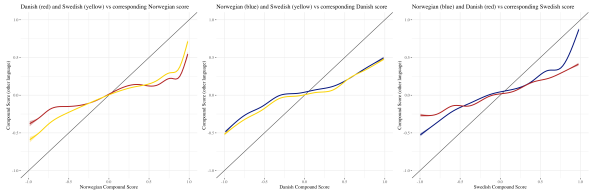

Classification of Uncertainty State
Resonance$_t$
Novelty$_t$
$\beta_1 = 0.53$
95%PI
95%CI


Classification of Uncertainty State
Resonance$_t$
Novelty$_t$
$\beta_1 = -0.11$
95%PI
95%CI


Classification of Uncertainty State
Resonance$_t$
Novelty$_t$
$\beta_1 = 0.15$
95%PI
95%CI

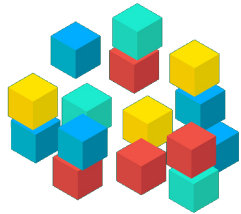

Classification of Uncertainty State
Resonance$_t$
Novelty$_t$
$\beta_1 = 0.6$
95%PI
95%CI


Moderna approved
Pfizer approved
Vaccinations begin
AstraZeneca (AZ) approved
AZ further halted
AZ withdrawn
AZ halted
Count of 'vaccin*' mentions
Compound Sentiment of 'vaccin*' mentions
2500
2000
1500
1000
500
0
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
15-12
01-01
15-01
01-02
15-02
01-03
15-03
01-04
15-04


Danish (red) and Swedish (yellow) vs corresponding Norwegian score
Norwegian (blue) and Swedish (yellow) vs corresponding Danish score
Norwegian (blue) and Danish (red) vs corresponding Swedish score
Compound Score (other language)
Norwegian Compound Score
Danish Compound Score
Swedish Compound Score
Shaded area indicates 95% CI

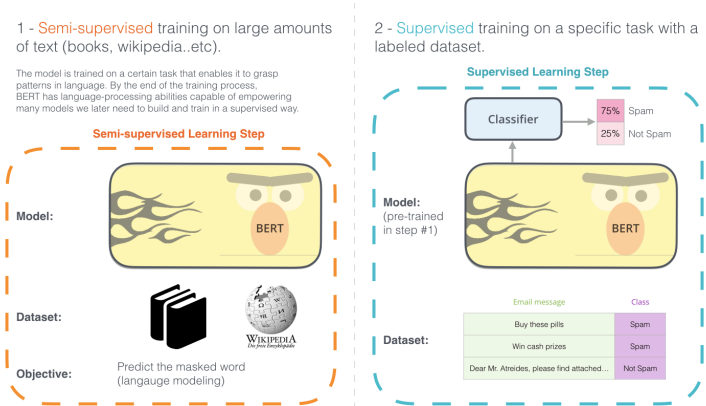
OUTLINE

1. During COVID-19
   - HOPE
   - FAIR ambition
   - Media monitoring
2. **Post COVID-19**
   - The unstructured challenge
   - Models cards and data sheets



The vast majority of HOPE's data are (potentially) sensitive and proprietary and cannot be shared in their original form. Can we find a way to convert those data to wel

Transformers and Transfer Learning



BERT broke several records for how well models can handle language tasks. Soon after the release of the paper describing the model, the team also open-sourced the model code and made it available for download versions of the model that were already pre-trained on massive datasets.

Danish Foundation Models

1. Develop and maintain **state-of-the-art** models for Danish
2. which are **well-validated** across a wide range of tasks
3. Furthermore, we wish to **ensure good documentation**, which allows users to assess the model for their use-case critically
4. **Open-source**, both model and source code

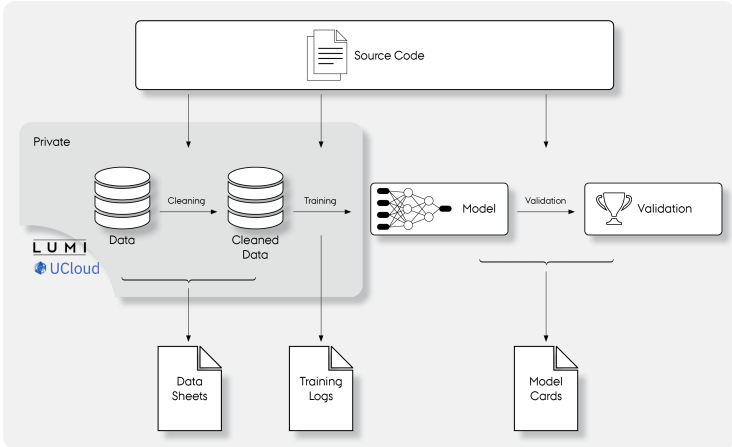
National Improvements

	Text		Audio	
	Danish	English	Danish	English
Architectures	BERT, (ELECTRA), (T5)	BERT, DeBERTaV3, GPT...	(Wav2vec2)	Wav2Vec, HuBERT, WavLM...
Largest pre-training corpus	1-1.6B Tokens	>100B tokens	1.300 hours	60.000 hours
Quality	Duplicate removal Removed cookies (Varied domains)	Near duplicate removal Quality filtered Varied domains Repetitious text removal	Single domain (audiobooks)	Audiobooks YouTube Podcasts
Number of parameters	14M - 100M (770M)	8M - 600B	135M	135M - 1B
Compute resource	20 days on 4 A100	Multiple months on >1024 TPUs w. Multiple of experimentation	7 days 4 titan GPUs	1 month on 8 A100 GPUs

Danish was lacking behind - 2021 the Royal Library of Norway releases a model that outperforms Danish models by 4-5 percentage points.

	Text		Audio	
	Danish	English	Danish	English
Architectures	BERT, (ELECTRA), (T5), DeBERTaV2	BERT, DeBERTaV3, GPT...	Wav2Vec2	Wav2Vec, HuBERT, WavLM...
Largest pre-training corpus	>100B Tokens	>100B tokens	140.000 hours	60.000 hours
Quality	Near duplicate removal Quality filtered Varied domains Repetitious text removal	Near duplicate removal Quality filtered Varied domains Repetitious text removal	Diverse radio shows Audiobooks	Audiobooks YouTube Podcasts
Number of parameters	14M - 300M (770M)	8M - 600B	300M	135M - 1B
Compute resource	20 days on 4 A100	Multiple months on >1024 TPUs w. Multiple of experimentation	8 days on 4 A100	1 month on 8 A100 GPUs

Not including multilingual models, ex. Whisper. Parenthesis indicates that we have the model, but it is small or trained on limited data.



Mainland Scandinavian NLU Benchmark: ScandEval

THANKS

kln@cas.au.dk
chc.au.dk

SLIDES

knielbo.github.io/files/kln-hope-os.pdf

ACKNOWLEDGEMENTS

This research was supported the “HOPE - How Democracies Cope with COVID-19”-project funded by The Carlsberg Foundation with grant CF20-0044, NeiC’s Nordic Digital Humanities Laboratory project, and DeiC Type-1 HPC with project DeiC-AU1-L-000001.