University of Antwerp

Evaluating Topic, Source and Viewpoint Diversity in Corona News

A Computational Analysis of Covid-19 News Coverage in the Netherlands

Elif Kılık

In a nutshell…

- **Leveraging generative large language models to extract information from news articles, which can be used to evaluate the diversity of the news content.**

Background of this study

- *"…although we witness a period of news abundance characterized by a proliferation of communication channels and outlets, the content of news is increasingly similar"* (Beckers et al., 2017, p.1666)
- Concerns about the impact on the quality of news (Saltzis, K., & Dickinson, 2008) and findings about decrease in diversity of news content (Hendrickx, & Van Remoortere, 2022)
- In the Netherlands → motion Peter Kwint (NL House of Representatives) – CvDM (Dutch Media Authority)

Definition of News (Content) Diversity

- *"… the heterogeneity of news content in terms of the plurality of actors, issues, and viewpoints (or frames) represented in the news."* (Beckers et al., 2017, p.1668)

Subdimensions of diversity in news content (Loecherbach et al., 2020):

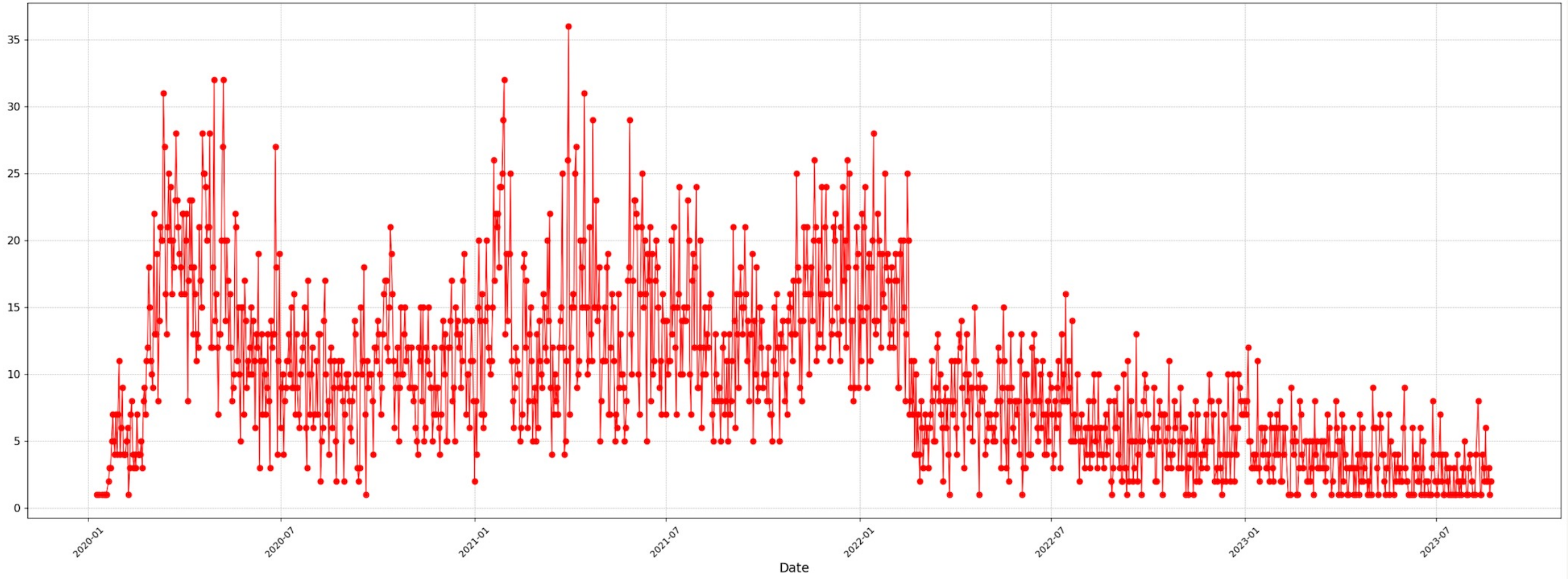
- **Topic (or issue) diversity:** the degree of variation in reporting on different topics issues
- **Actor (or source/entity) diversity:** the representation of (status positions of) sources used to create a news product
- **Diversity of viewpoints (perspectives):** the representation of ideas, perspectives, opinions

The Case Study – News about the Covid-19 Pandemic

- National and international extensive reporting over a limited period
  - Enough data for testing purposes
- Viewpoints on corona measures → allows more straightforward definition & measurement

Data

- 13586 news items retrieved from NOS.nl (Query: corona*, covid*, sars-cov-2*)

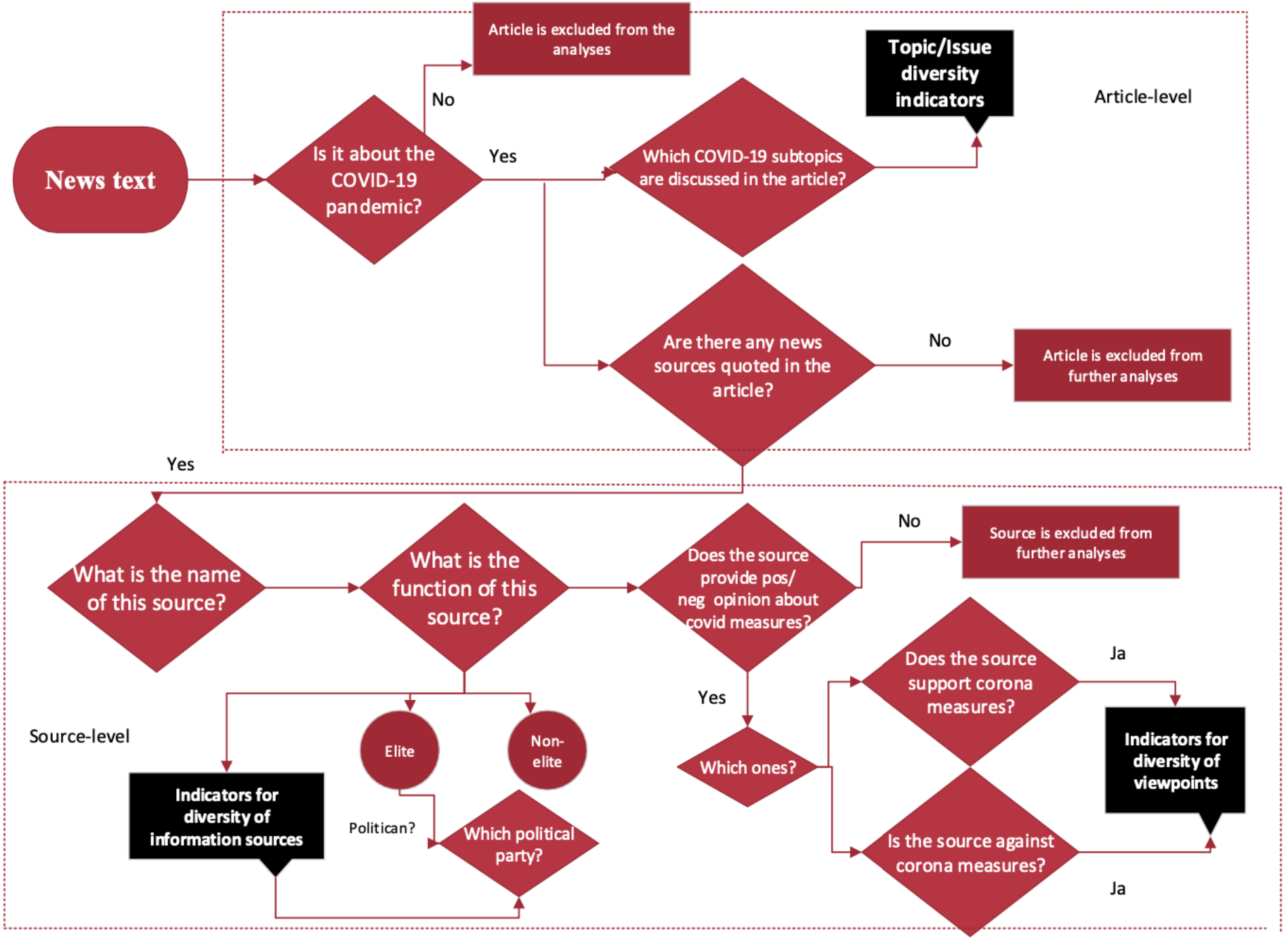


Methodology

- **Manual annotation**
  - 1000 randomly sampled articles, stratified based on date.
  - 472 already annotated by a trained student coder
  - This data is used as validation data for annotation with LLM's.

- **Annotation with Generative LLM's**
  - Tests (so far) with 3 Generative LLM's using 472 manually annotated articles:
    - meta-llama/Llama-2-7b-chat-hf
    - meta-llama/Llama-2-13b-chat-hf
    - berkeley-nest/Starling-LM-7B-alpha

Annotation Flow



University of Antwerp

Findings

(Work in progress…)

Finding 1: LLM's are able to detect main topic COVID-19

- **Classification of articles covid/non-covid**

- **Manual coding:**
  - *Variable 0: Is the article mainly about the corona pandemic?*

    This is a question to filter out articles that are not about the corona pandemic. The article can only be considered mainly about the corona pandemic if the majority of the article is about the pandemic or its (sub)topics (see V1 below). To be sure, you can check the word count of the parts dealing with the pandemic. If that is the majority, the article is mainly about the pandemic.

Finding 1: LLM's are able to detect main topic COVID-19

Input → article text, keywords and categories

Model Name	Examples in prompt	Kappa	Accuracy	Macro-averaged f1-score	N	Duration in mins
Starling 7b Alpha	Zero-shot	0.82	0.91	0.91	472	24
	One-shot	0.83	0.92	0.91	472	17
	Few-shot	**0.84**	**0.92**	**0.92**	**472**	**29**
Llama-2 7b Chat	**Zero-shot**	**0.73**	**0.86**	**0.86**	**472**	**15**
	One-shot	0.47	0.75	0.72	472	16
	Few-shot	0.71	0.86	0.85	419	41
Llama-2 13b Chat	Zero-shot	0.74	0.87	0.87	472	36
	One-shot	**0.80**	**0.90**	**0.90**	**472**	**23**
	Few-shot	0.78	0.89	0.89	426	59

Finding 2: Extracting subtopics is not straightforward

- **Pre-defined 14 topics (based on qualitative analysis of the news texts and previous studies):**
  a. Statistics (spread of the pandemic, number of positive cases, patients, fatalities
  b. Measures taken against the spread of the coronavirus pandemic
  c. COVID-19 tests and test procedures
  d. COVID-19 vaccines and vaccination procedures and campaigns
  e. Long COVID and long term effects of the coronavirus on health
  f. Impact of the COVID-19 and coronavirus pandemic on the healthcare system and medical response
  g. Scientific and medical knowledge on the coronavirus and COVID-19
  h. Economic impact of the coronavirus pandemic and recovery (industry & market, work culture, recovery packages)
  i. Social impact of the coronavirus pandemic (education, social gatherings, cultural events, sports)
  j. Impact of the COVID-19 and coronavirus pandemic on mental health
  k. Impact of the COVID-19 and coronavirus pandemic on individual rights and freedoms
  l. Misinformation and disinformation about the coronavirus, COVID-19 and/or the coronavirus pandemic
  m. Impact of the COVID-19 and coronavirus pandemic on politics and government (discussions in the Tweede Kamer, discussions on government's responsibility and ability)
  n. International discussions, cooperation and responses related to the coronavirus pandemic

Finding 2: Extracting subtopics is not straightforward

- **Some of these topics are not coded in the sample sufficient enough for the tests → 6 left**
  a. Statistics (spread of the pandemic, number of positive cases, patients, fatalities
  b. Measures taken against the spread of the coronavirus pandemic
  c. COVID-19 tests and test procedures
  d. COVID-19 vaccines and vaccination procedures and campaigns
  ~~e. Long COVID and long term effects of the coronavirus on health~~
  ~~f. Impact of the COVID-19 and coronavirus pandemic on the healthcare system and medical response~~
  ~~g. Scientific and medical knowledge on the coronavirus and COVID-19~~
  h. Economic impact of the coronavirus pandemic and recovery (industry & market, work culture, recovery packages)
  i. Social impact of the coronavirus pandemic (education, social gatherings, cultural events, sports)
  ~~j. Impact of the COVID-19 and coronavirus pandemic on mental health~~
  ~~k. Impact of the COVID-19 and coronavirus pandemic on individual rights and freedoms~~
  ~~l. Misinformation and disinformation about the coronavirus, COVID-19 and/or the coronavirus pandemic~~
  ~~m. Impact of the COVID-19 and coronavirus pandemic on politics and government (discussions in the Tweede Kamer, discussions on government's responsibility and ability)~~
  ~~n. International discussions, cooperation and responses related to the coronavirus pandemic~~

Finding 2: Extracting subtopics is not as straightforward

- **Binary classification of each label separately – so far the best working approach**

  You are a reliable assistant tasked with determining whether the news article below about the COVID-19 pandemic substantially discusses the subtopic: ”[insert subtopic name here]"

  Substantial discussion of a subtopic means that the article discusses one or more aspects of this subtopic in at least two sentences.

  The subtopic: ”[insert subtopic name here]" entails the following aspects:

  - .. List of what the subtopic discussions may contain

  Based on the information provided, determine if the article substantially discusses the subtopic: ”[insert subtopic name here]"

  Assign a value of 1 if the article substantially discusses the subtopic, and a value of 0 if the article does not substantially discuss the subtopic.

Finding 2: Model performances differ per subtopic

Statistics (spread of the pandemic, number of positive cases, patients, fatalities

→ Good results, but there are differences between models

Model Name	Examples in prompt	Kappa	Accuracy	Macro-averaged f1-score	N	Duration in mins
Starling 7b Alpha	Zero-shot	0.73	0.89	0.87	216	10
	One-shot	**0.81**	**0.92**	**0.90**	**216**	**10**
	Few-shot	0.78	0.90	0.89	216	11
Llama-2 7b Chat	Zero-shot	0.28	0.62	0.61	216	7
	One-shot	0.49	0.81	0.74	216	9
	Few-shot	**0.58**	**0.83**	**0.79**	**216**	**11**
Llama-2 13b Chat	**Zero-shot**	**0.50**	**0.78**	**0.75**	**216**	**12**
	One-shot	0.47	0.81	0.72	216	13
	Few-shot	0.50	0.81	0.74	216	16

Finding 2: Model performances differ per subtopic

Measures taken against the spread of the coronavirus pandemic

→ Has to be improved (models annotate too many cases of 1)

Model Name	Examples in prompt	Kappa	Accuracy	Macro-averaged f1-score	N	Duration in mins
Starling 7b Alpha	Zero-shot	0.38	0.73	0.69	216	11
	One-shot	0.43	0.75	0.71	216	10
	Few-shot	0.37	0.77	0.68	216	12
Llama-2 7b Chat	Zero-shot	0.05	0.38	0.36	216	7
	One-shot	0.28	0.69	0.64	216	9
	Few-shot	0.24	0.72	0.62	212	13
Llama-2 13b Chat	Zero-shot	0.28	0.62	0.61	216	12
	One-shot	0.24	0.67	0.62	216	13
	Few-shot	0.18	0.70	0.59	211	19

Finding 2: Model performances differ per subtopic

COVID-19 tests and test procedures

→ Has to be improved (models are not able to detect all relevant cases – low recall)

Model Name	Examples in prompt	Kappa	Accuracy	Macro-averaged f1-score	N	Duration in mins
Starling 7b Alpha	Zero-shot	0.34	0.89	0.66	216	10
	One-shot	0.46	0.89	0.73	216	10
	Few-shot	**0.47**	**0.89**	**0.74**	**216**	**11**
Llama-2 7b Chat	Zero-shot	0.10	0.40	0.39	216	7
	One-shot	**0.48**	**0.89**	**0.74**	**216**	**9**
	Few-shot	0.28	0.84	0.64	215	11
Llama-2 13b Chat	Zero-shot	0.27	0.67	0.59	216	14
	One-shot	0.21	0.88	0.59	216	13
	Few-shot	0.12	0.88	0.53	216	16

Finding 2: Model performances differ per subtopic

COVID-19 vaccines, vaccination procedures and campaigns

→ Good reliability for all the models

Model Name	Examples in prompt	Kappa	Accuracy	Macro-averaged f1-score	N	Duration in mins
Starling 7b Alpha	Zero-shot	0.73	0.92	0.86	216	10
	One-shot	**0.76**	**0.93**	**0.88**	**216**	**10**
	Few-shot	0.73	0.92	0.86	216	12
Llama-2 7b Chat	Zero-shot	0.58	0.83	0.78	216	7
	One-shot	**0.66**	**0.91**	**0.83**	**215**	**9**
	Few-shot	0.43	0.86	0.70	214	11
Llama-2 13b Chat	**Zero-shot**	**0.64**	**0.87**	**0.82**	**216**	**14**
	One-shot	0.52	0.87	0.76	216	13
	Few-shot	0.47	0.86	0.72	215	16

Finding 2: Model performances differ per subtopic

Economic impact of the COVID-19 pandemic and recovery

→ Easier to detect for all models

Model Name	Examples in prompt	Kappa	Accuracy	Macro-averaged f1-score	N	Duration in mins
Starling 7b Alpha	**Zero-shot**	**0.76**	**0.96**	**0.88**	**216**	**10**
	One-shot	0.73	0.95	0.86	216	9
	Few-shot	0.76	0.95	0.88	216	11
Llama-2 7b Chat	Zero-shot	0.12	0.48	0.43	216	7
	One-shot	0.62	0.94	0.81	216	8
	Few-shot	**0.66**	**0.95**	**0.83**	**216**	**10**
Llama-2 13b Chat	Zero-shot	0.27	0.71	0.59	216	17
	One-shot	0.65	0.94	0.83	216	12
	Few-shot	**0.82**	**0.97**	**0.91**	**216**	**13**

Finding 2: Model performances differ per subtopic

Social impact of the coronavirus pandemic

→ Anything and everything is social?

Model Name	Examples in prompt	Kappa	Accuracy	Macro-averaged f1-score	N	Duration in mins
Starling 7b Alpha	Zero-shot	0.26	0.75	0.62	216	11
	One-shot	0.16	0.54	0.50	216	10
	Few-shot	0.08	0.45	0.43	216	12
Llama-2 7b Chat	Zero-shot	0.03	0.46	0.41	214	7
	One-shot	0.10	0.78	0.55	216	10
	Few-shot	0.18	0.72	0.58	212	13
Llama-2 13b Chat	Zero-shot	0.07	0.54	0.47	216	14
	One-shot	0.00	0.68	0.49	216	14
	Few-shot	0.00	0.68	0.49	213	19

Next Steps

- Analysis of topics: where do we not agree? What kind of articles are difficult to label?
- Manual annotation of more articles → couldn't test everything
- Testing methods for extracting actors/sources from text (Named Entity Recognition)
- Matching actors to their functions and political parties
- Testing methods for viewpoint detection

- And for all these steps: all tips and suggestions are welcome ☺

Some info on testing with LLM's

- Refining prompts are very important but very tricky
- Getting the output format right takes a lot of trying (and failing)
- What works for one model may not work for another
- Giving examples to the model does not always improve the model's ability to detect information in the text
- Always validate, to make sure you have the right model for the task at hand, and the right prompt

- But, all in all the results are promising.

University of Antwerp

Questions?

Thank you for your attention!

Contact: elif.kilik@uantwerpen.be