CompText, Amsterdam, 4 May 2024

Few Shot Learning for Detecting Implicit Hypocrisy Accusations in Sustainability and Climate Debates

Paulina Garcia-Corral[1], Avishai Green[2], Hendrik Meyer[3], **Myrthe Reuver**[4], Xiaoyue Yan[5], Anke Stoll[6]

[1]Hertie School

[2] Hebrew University of Jerusalem

[3]University of Hamburg

[4]**Vrije Universiteit Amsterdam**

[5] University of Zurich

[6] Ilmenau University of Technology


#HACKICA23
TORONTO
This project: born at the ICA23 hackathon
NIKE
SPORTSWEA

Why hypocrisy? the Go-to Political Accusation

Hypocrisy **accusations** are abundant in politics

- Easy to make
- Effective
- In polarized polity - the only rhetorical tool available?

Image: Johan Eklund, Flickr

Politics: a “never ending fight to ferret out hypocrites” (Arendt, 2006, p. 93)

Climate - a prime hypocrisy discourse locus

- Meyer et al., 2023 & Falkenberg et al., 2022; Tyagi et al., 2021: Hypocrisy accusations are **central to increasingly polarized, cross-ideological online interactions** of climate change debates
- Luo et al., 2020: Climate action advocacy often framed as hypocritical by the opposition in newspaper articles
- Gunster et al., 2018: **Typology** of Climate Hypocrisy Accusations

Discourse type	**Political orientation**	**Goal**
Individual-lifestyle outrage	Right-wing	Outrage, shaming and silencing activists
Institutional cynicism	Right-wing	Fatalism
Call to action	Left-wing	Targeted anger, call to action
Reflexive	Left-wing	Productive discussions of how to facilitate change

Hypocrisy accusation: less attention in NLP tasks and low recall even with state-of-the-art models

Habernal et al. (2018): **sub-concept in fallacy detection**.

Instruction-tuned models (GPT and T-5) applied to climate change debate by Alhindi et al. (2022):

- In their 5 fallacy datasets, only **one dataset** with hypocrisy-related category, “whataboutism”
- Training on the other 4 datasets, they detect **"whataboutism" with .44 accuracy,**
- adding a definition leads to a small reduction to .43.

Piskorski et al. (2023) have designed a multilingual dataset on online news with an annotated hypocrisy accusation concept, as part of a **"persuasian techniques"** task.

- They also introduce an XLM-RoBERTa model as baseline. One of the debates: on climate change
- Their appendix reports a performance of the Whataboutism concept of .25% precision, with **extremely low recall (.034%)** leading to an F1 of .06.
- This concept is **only .05% of their dataset**

Hypocrisy Accusation Detection with small training samples

Detecting hypocrisy accusations in online debates (Reddit) with few examples.

Challenging because

- **Explicit vs. implicit**
  - *"Exactly! Imagine the US with three times the CO2 per capita to ask China to reduce emissions… THAT is hypocrisy."*
  - *Yet when I see those who make money on fossil fuels brag how clean they are ... seriously, how dare you?*

RQ1: Is **few-shot learning** suitable to detect accusations of hypocrisy?

RQ2: Can the few-shot learning model detect accusations of hypocrisy better than a BERT text classifier?

Fine-grained Reddit sustainability dataset

European Sustainable Initiatives Dataset (Reuver et. al, 2023)

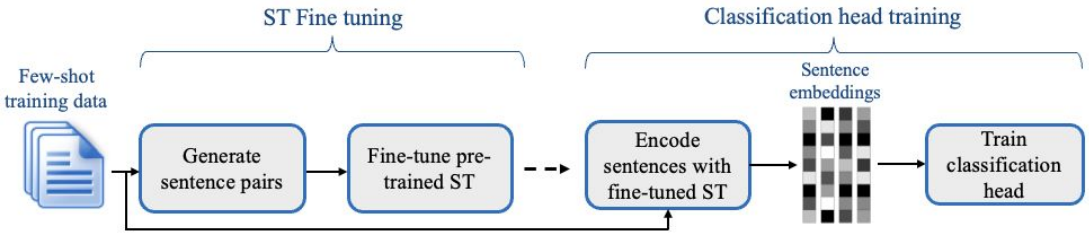
- **46.288 total** comments, on 3 reddit boards: eu, europe, and europes, with word2vec expanded wordlist
- 300 or **0.648%** hypocrisy mentions with regular expression
- contains: sub-discussions and interactions

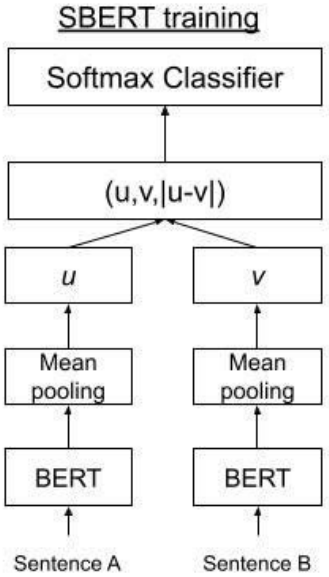
- Annotated during hackathon: quick 150 comments in debate contexts

Non-instruction-tuned few-shot method: SETFIT

- Triplet (text1, text2, label) **sentence embedding similarity training**; classification head on top
- ROBERTA-NLI in a setfit set-up.

- Results: with 10-25 shots similar to SVM (good for small data);
- around 75 to 100 shots greatly outperform SVM (.80 vs .65 accuracy)





Initial Results

- **SVM:** unsuccessful/incapable of predicting minority-class positive cases.
- **SETFIT** outperforms SVM when around 50-100 shots.
- **GPT 3.5 Turbo:** Most promising. 42/45 comments correctly classified!
  - including: *"Like Zuckerberg, who bought his neighbors houses to protect his privacy, while making billions selling other people privacy. ~~Hypocrisy is a virtue for these people~~"*

Conclusions:

- Task requires **precise conceptualization of the complex concept + careful evaluation**
- Few-shot classifying struggles with distinguishing implicit instances
- GPT looks promising…

Beyond the hackathon: more data? Using other instruction-tuned models?

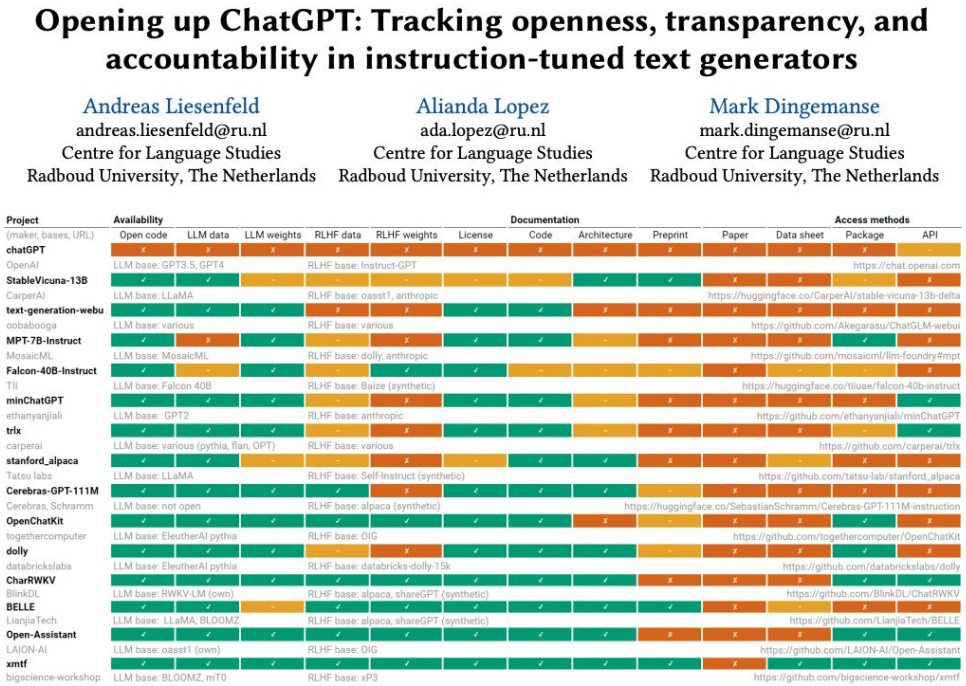
Mixtral-8x7B-Instruct: more open model than AI

- Mistral AI, French start-up.
- “Mixture of Experts”, like GPT (is rumored to be)

- Access through huggingface API access: free hourly rate limit

  Pro: Apache 2.0 license , free for academic and commercial usage, release papers about model development

- **Con:** They do release papers, better than OpenAI, not fully open, e.g. dataset information



Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators

Andreas Liesenfeld
andreas.liesenfeld@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

Alianda Lopez
ada.lopez@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

Mark Dingemanse
mark.dingemanse@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

Project	Availability						Documentation			Access methods			
(maker, bases, URL)	Open code	LLM data	LLM weights	RLHF data	RLHF weights	License	Code	Architecture	Preprint	Paper	Data sheet	Package	API
chatGPT	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	–
OpenAI	LLM base: GPT3.5, GPT4			RLHF base: Instruct-GPT									https://chat.openai.com
StableVicuna-13B	✓	✓	–	–	–	–	–	✓	✓	✗	✗	–	✗
CarperAI	LLM base: LLaMA			RLHF base: oasst1, anthropic									https://huggingface.co/CarperAI/stable-vicuna-13b-delta
text-generation-webu	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗
oobabooga	LLM base: various			RLHF base: various									https://github.com/Akegarasu/ChatGLM-webui
MPT-7B-Instruct	✓	✗	✓	–	✗	✓	✓	–	✗	✗	✗	✓	✗
MosaicML	LLM base: MosaicML			RLHF base: dolly, anthropic									https://github.com/mosaicml/llm-foundry#mpt
Falcon-40B-Instruct	✓	–	✓	–	✓	✓	–	–	–	✗	–	–	✗
TII	LLM base: Falcon 40B			RLHF base: Baize (synthetic)									https://huggingface.co/tiiuae/falcon-40b-instruct
minChatGPT	✓	✓	✓	–	✗	✓	✓	–	✗	✗	✗	✗	✓
ethanyanjiali	LLM base: GPT2			RLHF base: anthropic									https://github.com/ethanyanjiali/minChatGPT
trlx	✓	✓	✓	–	✗	✓	✓	–	✗	✗	✗	–	✓
carperai	LLM base: various (pythia, flan, OPT)			RLHF base: various									https://github.com/carperai/trlx
stanford_alpaca	✓	✓	–	–	✗	–	✓	✓	✗	✗	–	✗	✗
Tatsu labs	LLM base: LLaMA			RLHF base: Self-Instruct (synthetic)									https://github.com/tatsu-lab/stanford_alpaca
Cerebras-GPT-111M	✓	✓	✓	✓	✓	✓	✓	✓	–	✗	✗	✗	✗
Cerebras, Schramm	LLM base: not open			RLHF base: alpaca (synthetic)									https://huggingface.co/SebastianSchramm/Cerebras-GPT-111M-instruction
OpenChatKit	✓	✓	✓	✓	✓	✓	✓	✗	–	✗	✗	✓	✗
togethercomputer	LLM base: EleutherAI pythia			RLHF base: OIG									https://github.com/togethercomputer/OpenChatKit
dolly	✓	✓	✓	–	✗	✓	✓	✓	–	✗	✗	✓	✗
databrickslabs	LLM base: EleutherAI pythia			RLHF base: databricks-dolly-15k									https://github.com/databrickslabs/dolly
CharRWKV	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
BlinkDL	LLM base: RWKV-LM (own)			RLHF base: alpaca, shareGPT (synthetic)									https://github.com/BlinkDL/ChatRWKV
BELLE	✓	✓	–	✓	✓	✓	✓	✓	✓	✗	–	✗	✗
LianjiaTech	LLM base: LLaMA, BLOOMZ			RLHF base: alpaca, shareGPT (synthetic)									https://github.com/LianjiaTech/BELLE
Open-Assistant	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
LAION-AI	LLM base: oasst1 (own)			RLHF base: OIG									https://github.com/LAION-AI/Open-Assistant
xmtf	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
bigscience-workshop	LLM base: BLOOMZ, mT0			RLHF base: xP3									https://github.com/bigscience-workshop/xmtf

A Tale of Two Reddit Datasets

European Sustainable Initiatives (Reuver et. al, 2023)

- **46.288 total** comments, on 3 reddit boards: eu, europe, and europes, with word2vec expanded wordlist
- 300 or 0.648% hypocrisy mentions with regular expression
- contains: sub-discussions and interactions

Reddit Climate Change (Kaggle-released, large scrape dataset)

contains **all the posts and comments on Reddit** mentioning the terms "climate" and "change" until 2022-09-01.

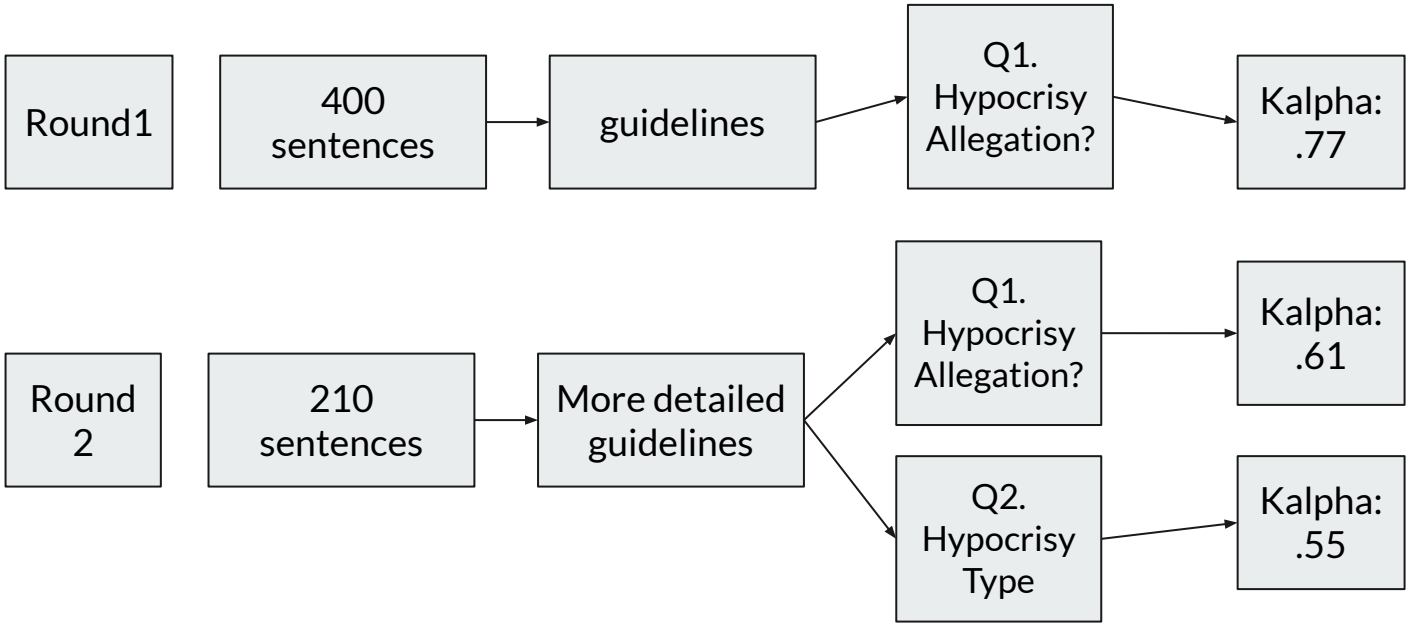
- **4.600.698** (!) comments in total
- **42.107** (!!) comments about hypocrisy (0.9%) with regular expression
- contains **no info** on sub-discussions and interactions, but much more data and subreddits.

1) regex pattern “hypocr*” 2) random sample

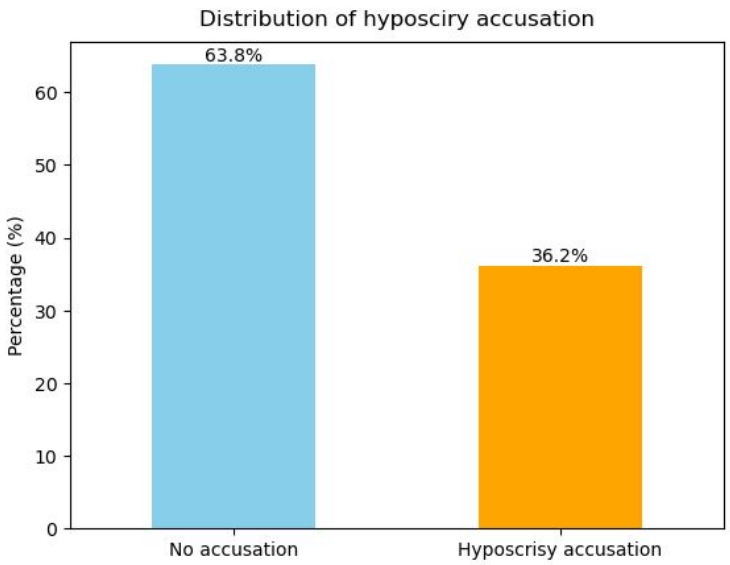
Merged into our final dataset → 300 dev examples, 500 test

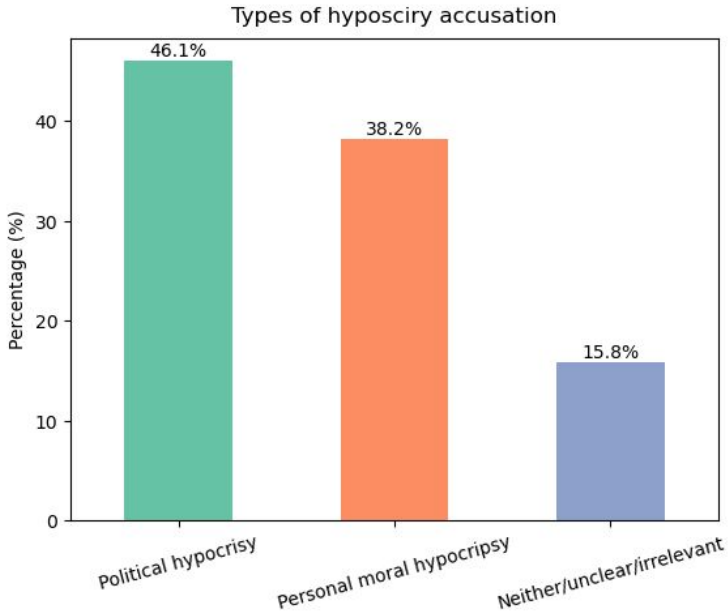
API: **Push-shift API access** ended in 2023. Both corpora were extracted before.

Annotation Process



Annotation process: development and test





Few-shot modelling: experimental setup

“ Annotate whether a comment contains a hypocrisy accusation.

Examples: comment_text, label

Comment text: {text}

Answer 1 for yes and 0 for no hypocrisy accusation:

Provide a reason:”

examples	“Provide reason”
0	yes
1	no
2	
3	

Experiments - how reliable is our model?

- Comparing **different prompt versions,** we found:

+ **“reason about it”:**

  Often does not lead to mistral adding reasons, but does seem to increase proficiency at task.

+ **Examples:**
  More is not always better

Possible future analyses

Two options, with our different corpora:

- **Going broad:** Analyzing **where most accusations in the big corpus** happen:
  - American politics boards vs European?
  - Certain years?
  - Related to certain issues/types of entities?

- Going **deep/fine-grained:**

  In the **European** dataset, we can see in *what depth of the interaction* the accusation happens, and which responses lead to it or come before. This information is not there for the broad Reddit dataset.

Conclusions

- Hypocrisy accusations in the climate context are **interesting,**
- These are **more textually complex** than initially thought, but received **less attention in current NLP and computational text analysis** work and datasets.
- **Data scarcity means few-shot learning methods** are an interesting road for developing models detecting such accusations, and initial results with instruction-tuned LLMs are interesting;
- Open(er) instruction-tuned LLMs can be a useful method for few-shot learning**, but still requires systematic analysis of prompt properties and outputs**


Don’t be a hypocrite…
ask a question
NIKE
SPORTSWEA