Towards Safer and Fairer Language Models: Mitigating Bias and Enhancing Regulatory Compliance

A Data Management Plan created using DMPonline.be

Creator: Pieter Delobelle

Affiliation: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Grant number / URL: PDMT2/23/047

ID: 205194

Start date: 08-12-2023

End date: 31-10-2024

Project abstract:

Language models such as GPT-3, GPT-4 and BERT are currently extremely popular and widely adopted by companies and individuals alike. With their ability to generate human-like responses and comprehend complex queries, these models have found their way into various applications, from chatbots like ChatGPT to content generation based on a user's input.

However, one of the major concerns with the use of language models is their susceptibility to biases against social groups. These biases can be introduced through the data used to train the models, as well as the algorithms and methods used to develop the models. Biases can have a significant impact, leading to unfair or discriminatory results, particularly for underrepresented groups. This problem is particularly understudied for languages other than English and outside of English-speaking countries, where applications of language models can lead to exacerbated biases.

Additionally, with the upcoming European AI Act, the need to better understand these limitations and biasesis becoming even more pressing.

Although some bias mitigation methods have been proposed, these methods are still lacking. Current approaches often rely on post-hoc debiasing techniques (i.e. after training) or on retraining a language model on `debiased' data, which can be computationally expensive and may not fully address the root causes of bias. As a result, there is a critical need for research that focuses on developing more robust and effective methods for mitigating bias in language models. By addressing this challenge, we can ensure that language models are more equitable and inclusive, and can be used to benefit all members of society, regardless of their background or identity.

Last modified: 07-03-2024

Towards Safer and Fairer Language Models: Mitigating Bias and Enhancing Regulatory Compliance

Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
		Indicate: ***N****(ew data) or* ***E****(xisting data)*	Indicate: **D**(igital) or **P**(hysical)	Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify)		Indicate: <1GB <100GB <1TB <5TB >5TB NA	
OSCAR	scraped data	E	D	T	irrelevant	>5 TB	
open subtitles	scraped data	E	D	T	irrelevant	<5 TB	
Code	source code	N	D	SO	repositories	<1GB	
LLMs	Language models	N and E	D	M	model weights	>5TB	
world-cup-2022-tweets	Labeled data	N	D	T	textual format (json, parquet)	<1GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

- models:
  - https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
  - https://huggingface.co/tiiuae/falcon-7b
- datasets:
  - https://huggingface.co/datasets/oscar-corpus/OSCAR-2301
  - https://huggingface.co/datasets/open_subtitles

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

- No

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or Data transfer agreements, Research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keepdata understandable and usable, for yourself and others, now and in the future (e.g. in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, codebook.tsv etc. where this information is recorded).

Created models will be documented with "Model cards".
Repositories will have a README.

Will a metadata standard be used to make it easier tofind and reuse the data?
If so, please specify which metadata standard will be used.

If not, please specify which metadata will be created to make the data easier to find and reuse.

- No

Data Storage & Back-up during the Research Project

Where will the data be stored?

- Large Volume Storage

All code will be stored on GitHub.

Trained models will be released on HuggingFace, as well as datasets. See:
- https://huggingface.co/datasets/pdelobelle/world-cup-2022-tweets

How will the data be backed up?

- Personal back-ups I make (specify below)
- Other (specify below)

Archived network storage is backed up. Code is saved in GitHub repositories as well.

Is there currently sufficient storage & backup capacity during the project?

If no or insufficient storage or backup capacities are available, explain how this will be taken care of.

- Yes

Both VSC and CS storage is not enough for saving all processed data during the project, so public datasets will be downloaded and deleted after processing.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

linux user groups.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

DTAI has its own large-scale (TB), long-term, NetApp storage system with backup. Costs are covered by the DTAI group.

Data Preservation after the end of the Research Project

Which data will be retained for 10 years (or longer, in agreement with other retention policies that are applicable) after the end of the project?

In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

- All data will be preserved for 10 years according to KU Leuven RDM policy

Where will these data be archived (stored and curated for the long-term)?

- Large Volume Storage (longterm for large volumes)

In the Dept CS central storage system (/cw/dtaiarch) on the NetApp storage system.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

DTAI has its own large-scale (TB), long-term, NetApp storage system with backup. Costs are covered by the DTAI group.

Data Sharing and Reuse

Will the data (or part of the data) be made available for reuse after/during the project?
Please explain per dataset or data type which data will be made available.

- Yes, as open data

If access is restricted, please specify who will be able to access the data and under what conditions.

Question not answered.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

Please explain per dataset or data type where appropriate.

Question not answered.

Where will the data be made available?

If already known, please provide a repository per dataset or data type.

- Other (specify below)

Standard websites such as huggingface and github. See:
- https://huggingface.co/datasets/pdelobelle/world-cup-2022-tweets

When will the data be made available?

- Other (specify below)

The latest upon publication, but we may choose to publish the data earlier.

Which data usage licenses are you going to provide?

If none, please explain why.

- MIT licence (code)

MIT is not only applicable to code, we also apply this licence to models that we will release.

Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.

- Yes, a PID will be added upon deposit in a data repository

What are the expected costs for data sharing? How will these costs be covered?

typically free.

Responsibilities

Who will manage data documentation and metadata during the research project?

Pieter Delobelle

Who will manage data storage and backup during the research project?

Pieter Delobelle

Who will manage data preservation and sharing?

Pieter Delobelle is responsible during the research project.
Luc De Raedt (PI) has the end responsibility and manages long term preservation and sharing.

Who will update and implement this DMP?

Pieter Delobelle is responsible during the research project.
Luc De Raedt (PI) has the end responsibility and manages long term preservation and sharing.