Detecting Harmful Content in Multilingual Raw Text

Language Models are revolutionizing NLP:

-Models like BERT have pushed state-of-the-art in many downstream tasks.

-Generative models like GPT-3 have delivered impressive results in few-shot and zero-shot learning.

-Lots of work being done in the understanding of Language Models.

Questions have been asked about the quality of the pre-training data:

-The data is automatically web-crawled.

-These models tend to repeat their pre-training data.

-Harmful content found on the web can be reproduced by these models!

The OSCAR Corpus:

-Contains more than 6TB or raw text across more than 150 different languages.

-The latest version contains metadata such as url, date of crawl and predicted languages for each entry.

-The latest version contains some annotations of potentially harmful or adult content.

We want to investigate the impact of harmful and noisy content in the pre-training data of recent Transformer-based Language Models

Project Goal

Assess the impact of potentially harmful content in pre-training data in recent Transformer-based Language Models. Propose new methods for annotating and filtering this content.

Involves

- data profiling/processing, train/test set creation
- model training, evaluation, selection and iteration

Learning Targets

- gain experience with large-scale data and state-of-the-art deep learning models
- gain work experience as Data Scientist

Requirements

- data wrangling skills, programming skills (Python!)
- relevant courses: Web Data Integration, Data Mining I & II, Text Analytics highly recommended

Organization: 4-6 people, 6 months, work as a complete team and in subgroups

Instructors: Pedro Ortiz Suarez

Related Literature:

[1] Kreutzer, Julia, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo et al. "Quality at a glance: An audit of web-crawled multilingual datasets." *arXiv preprint arXiv:2103.12028* (2021).

[2] Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610-623. 2021.

[3] Abadji, Julien, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. "Towards a Cleaner Document-Oriented Multilingual Crawled Corpus." *arXiv preprint arXiv:2201.06642* (2022).

[4] Caswell, Isaac, Theresa Breiner, Daan van Esch, and Ankur Bapna. "Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus." *arXiv preprint arXiv:2010.14571* (2020).