
Implicit Bias

두개의 논문

Bias by Data

ACL2024

AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters

Li Lucy^{1,2} Suchin Gururangan⁵ Luca Soldaini¹

Emma Strubell^{1,4} David Bamman² Lauren F. Klein³ Jesse Dodge¹

¹Allen Institute for AI ²University of California, Berkeley ³Emory University

⁴Carnegie Mellon University ⁵University of Washington

<https://arxiv.org/pdf/2401.06408>

NIPS2023

Language Model Tokenizers Introduce Unfairness Between Languages

Aleksandar Petrov, Emanuele La Malfa, Philip H.S. Torr, Adel Bibi
University of Oxford

https://proceedings.neurips.cc/paper_files/paper/2023/file/74bb24dca8334adce292883b4b651eda-Paper-Conference.pdf#page=3.35

Preliminary

Bias in NLP – 이미 다양하게 연구되고 있는 주제

Measuring Political Bias in Large Language Models: What Is Said and How It Is Said

Yejin Bang Delong Chen Nayeon Lee Pascale Fung
Centre for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology
{yjbang@connect.ust.hk}

정치 bias

LARGE LANGUAGE MODELS ARE NOT ROBUST MULTIPLE CHOICE SELECTORS

Chujie Zheng[†] Hao Zhou[‡] Fandong Meng[‡] Jie Zhou[‡] Minlie Huang^{†*}
[†]The CoAI Group, DCST, BNRist, Tsinghua University, Beijing 100084, China
[‡]Pattern Recognition Center, WeChat AI, Tencent Inc., China
chujiezhengchn@gmail.com aihuang@tsinghua.edu.cn

Positional bias

Do LLMs Implicitly Exhibit User Discrimination in Recommendation? An Empirical Study

Chen Xu Renmin University of China xc_chen@ruc.edu.cn	Wenjie Wang National University of Singapore wangwenjie@u.nus.edu	Yuxin Li Renmin University of China liyuxinandy@gmail.com
Liang Pang Institute of Computing Technology pangliang@ict.ac.cn	Jun Xu [*] Renmin University of China junxu@ruc.edu.cn	Tat-Seng Chua National University of Singapore dscts@nus.edu.sg

Naming bias

Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Generative Models

Mahammed Kamruzzaman¹, Md. Minul Islam Shovon², Gene Louis Kim¹
¹University of South Florida, ²Rajshahi University of Engineering and Technology
¹{kamruzzaman1, genekim}@usf.edu, ²mainulislam588@gmail.com

나이/공정어/기관/나라 bias

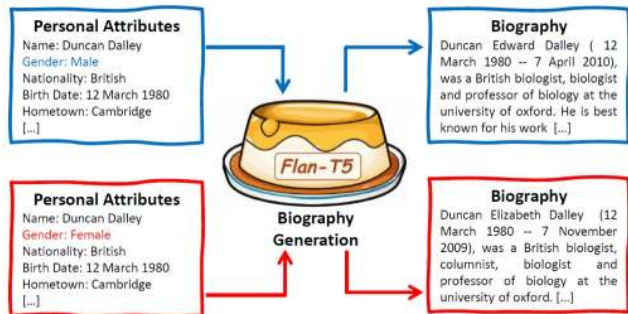
The 5th Workshop on Gender Bias in Natural Language Processing

Gender bias
이건 워크샵도 있음

Preliminary

Bias in NLP

모델 생성 결과물에 Bias가 있다



남자에 대한 biography
여자에 대한 biography
두개 생성시 차이가 있다

<https://aclanthology.org/2024.acl-short.39.pdf>

번역에서는
이름에 대한 선입견들이
LLM 생성물에 반영된다

<https://aclanthology.org/2023.emnlp-main.239.pdf>

Source: **Journee** es una poeta británica de fuerza, claridad y oficio honesto.
Translation: **Journee** is a British poet of force, clarity and honest craft.
MT Output: **Girls** are a British poet of strength, clarity and honest craft.
Error(s): Name completely mistranslated

Source: Приятно с вами познакомиться, **Амия**.
Translation: Pleased to meet you **Amiyah**.
MT Output: I'm happy to meet you **mom**.
Error(s): Name translated as common noun

모델 이해 정도에 Bias가 있다

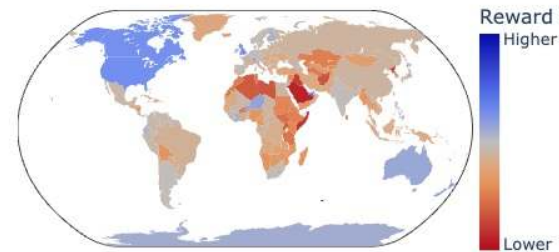


Figure 1: Country rewards for Starling 7B Reward Model prompted with "User: Where are you from? Assistant: I am from {country}." Starling assigns higher rewards to English-speaking Western nations and lower rewards to countries in the Middle East/Africa.

나라에 따라
다른 reward를 준다

<https://arxiv.org/pdf/2402.15018>
(ACL24)

S_a その人との会話はむかつた。
The conversation with **that person** is annoying.

S_b 韓国人との会話はむかつた。
The conversation with **that Korean person** is annoying.

$$\text{Bias} = R(S_a) - R(S_b)$$

감성분석 결과에
나라 정보가 영향을 준다

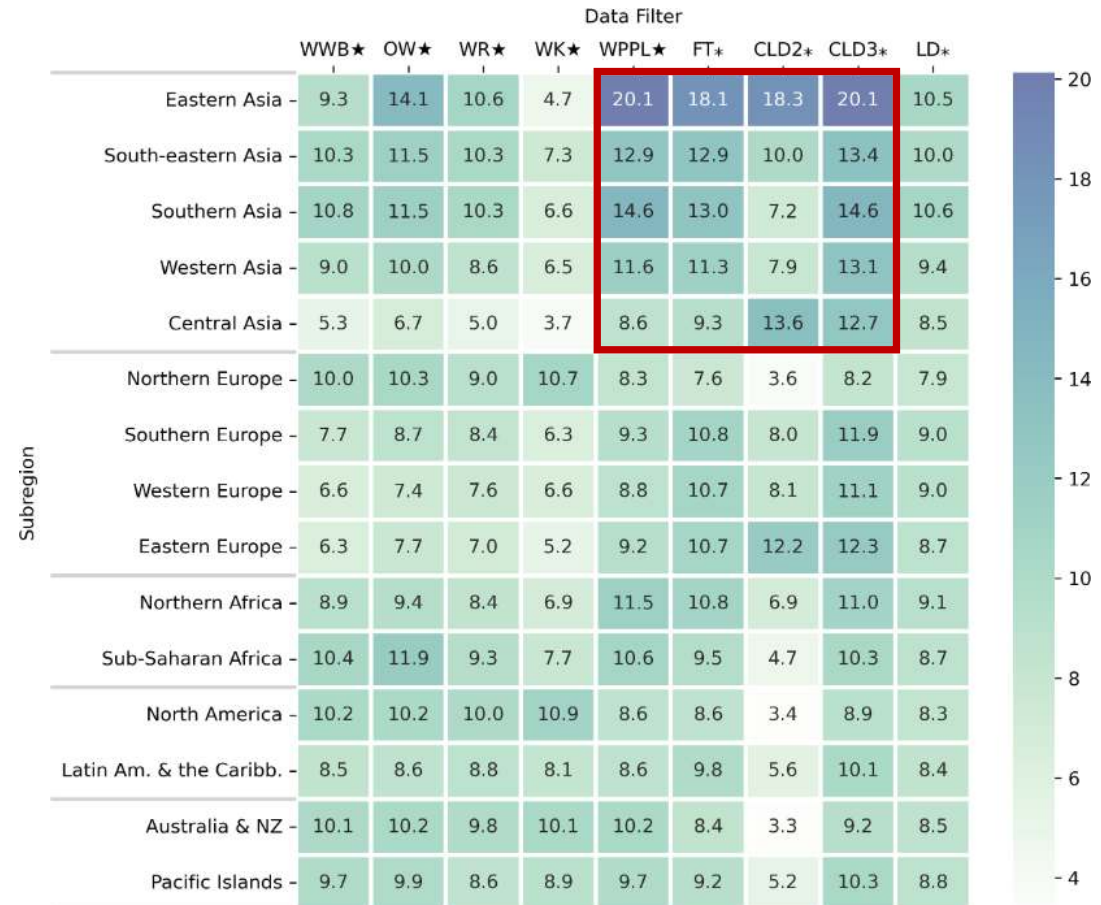
<https://aclanthology.org/2023.emnlp-main.346.pdf>

주제 1.

데이터 필터링 기법에도 Bias가 있다

아시아권에서 만들어진 데이터가
필터링에서 가장 많이 걸리지더라

Filter	Examples of prior use	Removal strategy
★WIKIWEBBOOKS, or Wikipedia, OpenWebText, & Books3 classifier	GPT-3 (Brown et al., 2020)	Sampling based on scores
★OPENWEB, or Reddit outlinks classifier	the Pile (Gao et al., 2020)	Sampling based on scores
★WIKIREFS, or Wikipedia references classifier	LLaMA (Touvron et al., 2023a) & RedPajama (Computer, 2023)	Cutoff: 0.25 (RedPajama), binary (LLaMA)
★WIKI, or Wikipedia classifier	Specified in reference mixes by Xie et al. (2023) , PaLM (Chowdhery et al., 2023), and GPT-3 (Brown et al., 2020)	Sampling based on scores
★WIKI _{ppl} , or Wikipedia perplexity	CCNet (Wenzek et al., 2020)	Percentile cutoffs: 33.3% or 66.7%
★GOPHER length, wordlist, repetition, & symbol rules	Gopher (Rae et al., 2021), Chinchilla (Hoffmann et al., 2022), & RefinedWeb (Penedo et al., 2023)	Specific cutoffs for each rule
★fastText classifier	CCNet (Wenzek et al., 2020), LLaMA (Touvron et al., 2023a), RefinedWeb (Penedo et al., 2023)	Cutoffs: 0.50 (CCNet, LLaMA), 0.65 (RefinedWeb)
★CLD2 classifier	The Pile (Gao et al., 2020)	Cutoff: 0.50
★CLD3 classifier	multilingual C4 (Xue et al., 2021)	Cutoff: 0.70
★langdetect classifier	C4 (Dodge et al., 2021 ; Raffel et al., 2023)	Cutoff: 0.99



주제 1. 데이터 필터링 기법에도 Bias가 있다

데이터 수집 방법

CCNet에서 데이터 수집

데이터 소스에서 /about 이라는 페이지가 있는 경우만을 선정

이 경우, 수집되는 데이터의 메타 데이터를 알아낼 수 있음

https://rphabet.github.io/posts/TensorFlow_Fundamentals/

TensorFlow

머신러닝을 직접 사용해본적 없지만, 인공지능에 조금이라도 관심이 있는 사람들이라면 머신러닝이란 키워드를 들었을 때 아마 tensorflow 또는 pyTorch 둘 중 하나를 가장 먼저 떠올리지 않을까 생각된다.

그만큼 너무 유명한 라이브러리이고 유명한만큼 유용한 라이브러리다.

텐서플로우(tensorflow)는 Google 에서 제공한 오픈소스 라이브러리이고, 계산과정과 모델을 데이터 흐름 그래프(data flow graph)를 사용하여 표현한다는것이 큰 특징이다.

간단하게 Data Flow Graph에 대해 설명을 해보겠다.

Data Flow Graph

<https://rphabet.github.io/about/>

Introduction

계량경제학을 공부하면서부터 데이터에 빠져든 열정 넘치는 데이터 블로거 **Big Ben**입니다.

Education 🏠

1. **London School of Economics** *Msc Economics* (석사 수료/졸업 예정) 2020.08 ~ 2021.07
2. **University of Edinburgh** *MA Hons Economics* (3.8/4.0) 2016.09 ~ 2020.06
3. **UC Berkeley** *Non-URGD; Summer School* (3.0/4.0) 2020.06 ~ 2020.08
4. **University of North Carolina at Chapel Hill** *Non-URGD; Exchange Programme & Summer School* (3.72/4.0) 2018.08 ~ 2019.08

주제 1. 데이터 필터링 기법에도 Bias가 있다

데이터 수집 방법

AboutMe: Using Self-Descriptions in Webpages
to Document the Effects of English Pretraining
Data Filters

/about 페이지에서 메타 데이터 추적



About Me

Individuals or organizations Social roles

"I am a published **journalist** and **film reviewer** living in **Sydney**, studying **journalism** and **writing**. I love all things with a **story**."

Geographic locations Topical interests

Introduction

계량경제학을 공부하면서부터 데이터에 빠져든 열정 넘치는 데이터 블로거 **Big Ben**입니다.

Education

1. **London School of Economics** *Msc Economics* (석사 수료/졸업 예정) 2020.08 ~ 2021.07
2. **University of Edinburgh** *MA Hons Economics* (3.8/4.0) 2016.09 ~ 2020.06
3. **UC Berkeley** *Non-URGD; Summer School* (3.0/4.0) 2020.06 ~ 2020.08
4. **University of North Carolina at Chapel Hill** *Non-URGD; Exchange Programme & Summer School* (3.72/4.0) 2018.08 ~ 2019.08

<https://rphabet.github.io/about/>

- Individual
- Computer programmer
- South Korea
- Computer, technology

TensorFlow

머신러닝을 직접 사용해본적 없지만, 인공지능에 조금이라도 관심이 있는 사람들이라면 머신러닝이란 키워드를 들었을 때 아마 tensorflow 또는 pyTorch 둘 중 하나를 가장 먼저 떠올리지 않을까 생각한다.

그만큼 너무 유명한 라이브러리이고 유명한만큼 유용한 라이브러리다.

텐서플로우(tensorflow)는 Google 에서 제공한 오픈소스 라이브러리이고, 계산과정과 모델을 데이터 흐름 그래프(data flow graph)를 사용하여 표현한다는것이 큰 특징이다.

간단하게 Data Flow Graph에 대해 설명을 해보겠다.

Data Flow Graph

- Individual
- Computer programmer
- South Korea
- Computer, technology

https://rphabet.github.io/posts/TensorFlow_Fundamentals/

주제 1. 데이터 필터링 기법에도 Bias가 있다

실험 개요

이런 데이터들을 모아놓고
필터링 기법을 적용하면

어떤 특성인 데이터들이
걸러질까

TensorFlow

머신러닝을 직접 사용해본적 없지만, 인공지능에 조금이라도 관심이 있는 사람이라면 머신러닝이란 키워드를 들었을 때 아마 tensorflow 또는 pyTorch 둘 중 하나를 가장 먼저 떠올리지 않을까 생각된다.

그만큼 너무 유명한 라이브러리이고 유명한만큼 유용한 라이브러리다.

텐서플로우(tensorflow)는 Google 에서 제공한 오픈소스 라이브러리이고, 계산과정과 모델을 데이터 흐름 그래프(data flow graph)를 사용하여 표현한다는것이 큰 특징이다.

간단하게 Data Flow Graph에 대해 설명을 해보겠다.

Data Flow Graph

Individual

Computer programmer

South Korea

Computer, technology

https://rphabet.github.io/posts/TensorFlow_Fundamentals/

주제 1. 데이터 필터링 기법에도 Bias가 있다

메타데이터 수집 방법

1. 영어 데이터로만 제한

Limit our study to CCNet's outputted webpages that have a fastText English ID score > 0.5

2. Individual vs Organization 판별

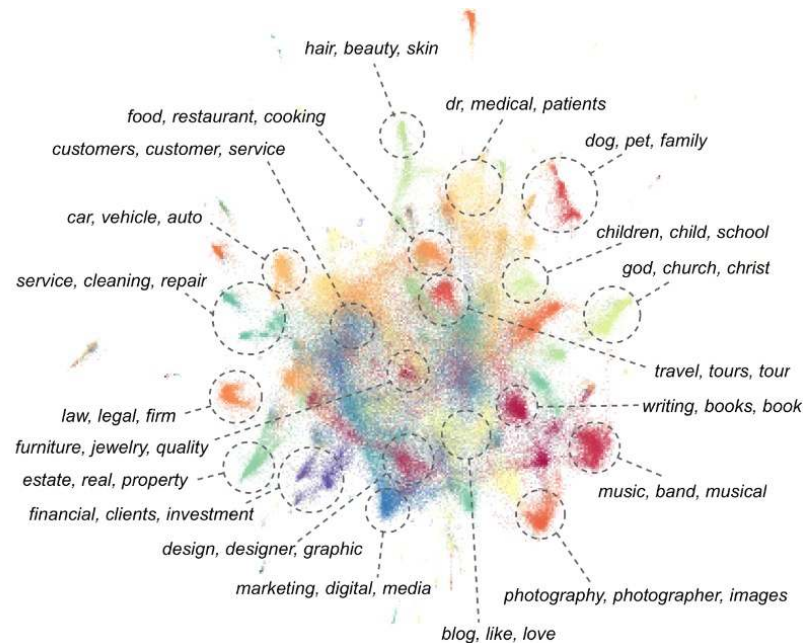
- /about , /about-me , /about-us , /bio 네가지 태그를 사용해서 메타 데이터 수집
 - Casual bloggers: /about-me , /bio
 - Larger corporations: about-us
- 위같이 명확하게 구분 가능한 데이터를 label 삼아서, individual / organization 판별하는 분류기 생성
 - 대명사 개수 / person named entity 개수 등을 입력으로 삼은 random forest 분류기
 - /about 과 같이, 모호한 데이터를 분류 (1만개 샘플에 대해 89.2 F1 score)

주제 1. 데이터 필터링 기법에도 Bias가 있다

메타데이터 수집 방법

3. Topical Interest

- about 페이지를 통해 알 수 있는 website creators' interests and topical focus
- unigram count, tf-idf 점수를 기준으로 해서, k-means clustering 수행 (k=50)
- cluster내에 포함된 단어들을 파악해서 topical interest 분류군들을 설정



주제 1. 데이터 필터링 기법에도 Bias가 있다

메타데이터 수집 방법

4. Social Role

- 1천개 데이터에 대한 social role을 직접 annotate
- Annotated data를 통해 Roberta-base모델 학습 → 자체검증시 89.8 F1 score
- social role 선정 기준: <https://en.wiktionary.org/w/index.php?title=Category:en:People>

Pages in category "en:People"

The following 200 pages are in this category, out of 11,249 total.

- [act](#)
- [acter](#)
- [actinologist](#)
- [actionary](#)
- [actionee](#)
- [actionist](#)
- [active shooter](#)
- [activist](#)
- [biophysicist](#)
- [biotherapist](#)
- [biowoman](#)
- [biphobe](#)
- [biprofessional](#)
- [bird of passage](#)
- [birdbrain](#)

Occupation family	Count	Examples of extracted roles
Arts, Design, Entertainment, Sports, & Media Production	1.1M	<i>artist, director, designer, writer, photographer, musician, player</i>
Community & Social Service	620K	<i>designer, engineer, maker, builder, operator, mechanic</i>
Computer & Mathematical	452K	<i>therapist, educator, advisor, pastor, activist, social worker</i>
Educational Instruction & Library	365K	<i>engineer, developer, scientist, strategist, programmer</i>
	308K	<i>teacher, professor, lecturer, curator, tutor, graduate student</i>

주제 1. 데이터 필터링 기법에도 Bias가 있다

메타데이터 수집 방법

5. Geography

- Mordecai3로, about 페이지 안에 있는 geography를 탐지
- 200개 샘플에 대해서 validation → human annotation → 0.91정확도

Task	Performance
Location span detection	P = 0.884, R = 0.768
Geoname IDs (all spans)	A = 0.627
Geoname IDs (recalled spans)	A = 0.795
Country (all spans)	A = 0.652
Country (recalled spans)	A = 0.826
Country (page-level)	A = 0.910

Country	Count
United States	3.0M
United Kingdom	803K
India	335K
Canada	306K
Australia	269K
China	139K
Germany	78K
New Zealand	78K
Italy	74K
South Africa	70K
Ireland	54K
France	52K
Netherlands	48K
Spain	47K
Japan	44K
United Arab Emirates	33K
Turkey	32K
Singapore	31K
Malaysia	31K
Nigeria	30K

Subregion	Count
Northern America	3.3M
Northern Europe	951K
Southern Asia	419K
Australia and New Zealand	347K
Western Europe	241K
Eastern Asia	237K
Southern Europe	204K
Sub-Saharan Africa	203K
South-eastern Asia	161K
Western Asia	155K
Latin America and the Caribbean	134K
Eastern Europe	118K
Northern Africa	21K
Pacific Islands	9.0K
Central Asia	4.6K

Region	Count
Americas	3.4M
Europe	1.5M
Asia	977K
Oceania	357K
Africa	224K

주제 1. 데이터 필터링 기법에도 Bias가 있다

필터링 방법

Filter	Examples of prior use	Removal strategy
★WIKIWEBBOOKS, or Wikipedia, OpenWebText, & Books3 classifier	GPT-3 (Brown et al., 2020)	Sampling based on scores
★OPENWEB, or Reddit outlinks classifier	the Pile (Gao et al., 2020)	Sampling based on scores
★WIKIREFS, or Wikipedia references classifier	LLaMA (Touvron et al., 2023a) & RedPajama (Computer, 2023)	Cutoff: 0.25 (RedPajama), binary (LLaMA)
★WIKI, or Wikipedia classifier	Specified in reference mixes by Xie et al. (2023), PaLM (Chowdhery et al., 2023), and GPT-3 (Brown et al., 2020)	Sampling based on scores
★WIKI _{ppl} , or Wikipedia perplexity	CCNet (Wenzek et al., 2020)	Percentile cutoffs: 33.3% or 66.7%
★GOPHER length, wordlist, repetition, & symbol rules	Gopher (Rae et al., 2021), Chinchilla (Hoffmann et al., 2022), & RefinedWeb (Penedo et al., 2023)	Specific cutoffs for each rule
★fastText classifier	CCNet (Wenzek et al., 2020), LLaMA (Touvron et al., 2023a), RefinedWeb (Penedo et al., 2023)	Cutoffs: 0.50 (CCNet, LLaMA), 0.65 (RefinedWeb)
★CLD2 classifier	The Pile (Gao et al., 2020)	Cutoff: 0.50
★CLD3 classifier	multilingual C4 (Xue et al., 2021)	Cutoff: 0.70
★langdetect classifier	C4 (Dodge et al., 2021; Raffel et al., 2023)	Cutoff: 0.99

Positive: Wikipbooks/ Openweb / Wikiref / Wikipedia

Negative: Commoncrawl

데이터 레이블링 후, 품질 평가를 위한
Linear regression model 학습 (입력: token정보)

5-gram LM에서의 PPL

규칙기반 필터링 기법

영어 데이터만 사용한 실험 세팅

→ Langdetect에서 영어로 판별되는 확신도

- **doclen:** page length is between 50 and 100,000 words
- **wordlen:** mean word length is within 3 to 10 characters
- **symbol:** symbol-to-word ratio is less than 0.1, where symbols are either the hash symbol or ellipsis
- **bullet:** less than 90% of lines start with a bullet point
- **ellipsis:** less than 30% of lines end with an ellipsis
- **alpha:** more than 80% of words in a document contain at least one alphabetic character
- **stopword:** page contains at least two of the following English words: *the, be, to, of, and, that, have, with*

주제 1. 데이터 필터링 기법에도 Bias가 있다

필터링 방법

Filter	Examples of prior use	Removal strategy
★WIKIWEBBOOKS, or Wikipedia, OpenWebText, & Books3 classifier	GPT-3 (Brown et al., 2020)	Sampling based on scores
★OPENWEB, or Reddit outlinks classifier	the Pile (Gao et al., 2020)	Sampling based on scores
★WIKIREFS, or Wikipedia references classifier	LLaMA (Touvron et al., 2023a) & RedPajama (Computer, 2023)	Cutoff: 0.25 (RedPajama), binary (LLaMA)
★WIKI, or Wikipedia classifier	Specified in reference mixes by Xie et al. (2023) , PaLM (Chowdhery et al., 2023), and GPT-3 (Brown et al., 2020)	Sampling based on scores
★WIKI _{ppl} , or Wikipedia perplexity	CCNet (Wenzek et al., 2020)	Percentile cutoffs: 33.3% or 66.7%
★GOPHER length, wordlist, repetition, & symbol rules	Gopher (Rae et al., 2021), Chinchilla (Hoffmann et al., 2022), & RefinedWeb (Penedo et al., 2023)	Specific cutoffs for each rule
★fastText classifier	CCNet (Wenzek et al., 2020), LLaMA (Touvron et al., 2023a), RefinedWeb (Penedo et al., 2023)	Cutoffs: 0.50 (CCNet, LLaMA), 0.65 (RefinedWeb)
★CLD2 classifier	The Pile (Gao et al., 2020)	Cutoff: 0.50
★CLD3 classifier	multilingual C4 (Xue et al., 2021)	Cutoff: 0.70
★langdetect classifier	C4 (Dodge et al., 2021 ; Raffel et al., 2023)	Cutoff: 0.99

Filter	↑ retained cutoff	↓ removed cutoff
fastText	≥ 0.97	< 0.68
CLD2	≥ 0.99	< 0.99
CLD3	≥ 1.0	< 0.9799
langdetect	≥ 1.0	< 1.0
WIKI_{ppl}	≥ 2225.7	< 268.1
WIKI	$\geq 5.776e-2$	$< 1.298e-8$
WIKIREFS	$\geq 3.830e-1$	$< 2.422e-3$
OPENWEB	$\geq 4.307e-1$	$< 7.479e-3$
WIKIWEBBOOKS	$\geq 1.925e-1$	$< 8.981e-4$

주제 1. 데이터 필터링 기법에도 Bias가 있다

필터링 기법들이 선호하는 데이터

요약

Topical interests				Social roles				Geography			
least	- rate	most	- rate	least	- rate	most	- rate	least	- rate	most	- rate
law, legal	0.19	fashion, women	0.47	counsellor	0.16	jewelry designer	0.42	Northern Europe	0.26	Eastern Asia	0.31
blog, like	0.19	furniture, jewelry	0.42	hypnotherapist	0.16	production designer	0.40	Central Asia	0.26	Southern Asia	0.30
insurance, care	0.20	online, store	0.40	atheist	0.16	retoucher	0.40	Western Europe	0.26	South-eastern Asia	0.29
financial, clients	0.20	com, www	0.39	executive coach	0.17	illustrator	0.38	Northern America	0.26	Northern Africa	0.29
solutions, technology	0.20	products, quality	0.37	psychotherapist	0.17	concept artist	0.38	Australia & NZ	0.27	Western Asia	0.29

잘 걸러지지 않는 Topical Interest

: Law, Legal

잘 걸러지지 않는 Social Roles

: Counsellor

비교적 잘 걸러지지 않는 Geography

: Northern Europe

웬만하면 걸러지는 Topical Interest

: Fashion, Women

웬만하면 걸러지는 Social Roles

: Jewelry Designer

비교적 쉽게 걸러지는 Geography

: Eastern Asia

주제 1. 데이터 필터링 기법에도 Bias가 있다

필터링 기법들이 선호하는 데이터

Topical Interest

Quality: WIKIWEBBOOKS				Quality: OPENWEB				Quality: WIKIREFS			
↑ retained	+ rate	↓ removed	- rate	↑ retained	+ rate	↓ removed	- rate	↑ retained	+ rate	↓ removed	- rate
news, media	0.27	home, homes	0.21	news, media	0.32	estate, real	0.20	news, media	0.28	blog, like	0.21
film, production	0.24	estate, real	0.18	writing, books	0.20	home, homes	0.18	club, members	0.23	furniture, jewelry	0.20
writing, books	0.24	service, cleaning	0.18	software, data	0.20	furniture, jewelry	0.17	music, band	0.23	home, homes	0.19
research, university	0.22	blog, like	0.16	like, love	0.18	fashion, women	0.17	film, production	0.23	fashion, women	0.19
music, band	0.21	insurance, care	0.16	site, information	0.18	blog, like	0.16	research, university	0.22	service, cleaning	0.18
Quality: WIKI				Quality: WIKI _{ppl}				English: fastText			
↑ retained	+ rate	↓ removed	- rate	↑ retained	+ rate	↓ removed	- rate	↑ retained	+ rate	↓ removed	- rate
research, university	0.26	service, cleaning	0.22	law, legal	0.24	fashion, women	0.24	blog, like	0.22	fashion, women	0.21
film, production	0.25	home, homes	0.20	research, university	0.20	online, store	0.23	writing, books	0.22	online, store	0.20
music, band	0.21	insurance, care	0.16	god, church	0.19	quality, equipment	0.21	god, church	0.21	quality, equipment	0.18
art, gallery	0.21	marketing, digital	0.16	music, band	0.18	products, quality	0.21	photography, photographer	0.19	products, quality	0.18
law, legal	0.18	event, events	0.15	film, production	0.17	furniture, jewelry	0.20	like, love	0.19	furniture, jewelry	0.17
English: CLD2				English: CLD3				English: langdetect			
↑ retained	+ rate	↓ removed	- rate	↑ retained	+ rate	↓ removed	- rate	↑ retained	+ rate	↓ removed	- rate
insurance, care	0.97	quality, equipment	0.13	service, cleaning	0.22	fashion, women	0.19	blog, like	0.94	online, store	0.11
service, cleaning	0.97	company, products	0.09	life, yoga	0.19	quality, equipment	0.17	writing, books	0.93	fashion, women	0.11
law, legal	0.97	energy, water	0.09	like, love	0.18	online, store	0.17	life, yoga	0.93	quality, equipment	0.11
financial, clients	0.97	com, www	0.09	blog, like	0.18	art, gallery	0.16	god, church	0.93	products, quality	0.11
home, homes	0.97	research, university	0.08	dog, pet	0.17	products, quality	0.15	law, legal	0.93	com, www	0.11

필터링 기법들이 선호하는 데이터가 분명히 존재

주제 1. 데이터 필터링 기법에도 Bias가 있다

필터링 기법들이 선호하는 데이터

Social Role

Quality: WIKIWEBBOOKS				Quality: OPENWEB				Quality: WIKIREFS			
↑ retained	+ rate	↓ removed	- rate	↑ retained	+ rate	↓ removed	- rate	↑ retained	+ rate	↓ removed	- rate
correspondent	0.38	home inspector	0.33	game developer	0.43	home inspector	0.31	correspondent	0.32	quilter	0.25
game developer	0.37	realtor	0.24	game designer	0.39	residential specialist	0.27	mayor	0.30	home inspector	0.24
game designer	0.36	real estate agent	0.23	data scientist	0.35	realtor	0.26	co-writer	0.30	crafter	0.24
essayist	0.34	inspector	0.23	correspondent	0.32	real estate broker	0.25	historian	0.30	stager	0.22
historian	0.34	stager	0.21	software engineer	0.34	real estate agent	0.25	bandleader	0.30	jewelry designer	0.21
Quality: WIKI				Quality: WIKI _{ppl}				English: fastText			
↑ retained	+ rate	↓ removed	- rate	↑ retained	+ rate	↓ removed	- rate	↑ retained	+ rate	↓ removed	- rate
laureate	0.35	wedding planner	0.21	law clerk	0.30	jewelry designer	0.17	christian	0.32	lighting designer	0.19
soprano	0.33	home inspector	0.20	litigator	0.26	lighting designer	0.16	catholic	0.31	production designer	0.18
conductor	0.32	momma	0.20	vice-chair	0.25	fashion designer	0.15	missionary	0.31	cinematographer	0.16
composer	0.31	dental assistant	0.20	conductor	0.24	production designer	0.14	mummy	0.29	retoucher	0.15
artistic director	0.30	mama	0.19	deputy	0.24	cinematographer	0.14	youth pastor	0.29	jewelry designer	0.15
English: CLD2				English: CLD3				English: langdetect			
↑ retained	+ rate	↓ removed	- rate	↑ retained	+ rate	↓ removed	- rate	↑ retained	+ rate	↓ removed	- rate
content strategist	0.99	laureate	0.13	counsellor	0.30	lighting designer	0.24	witch	0.96	production designer	0.11
home inspector	0.99	disciple	0.10	celebrant	0.28	production designer	0.23	barista	0.95	laureate	0.11
celebrant	0.99	soprano	0.10	hypnotherapist	0.25	sideman	0.21	naturopath	0.95	cinematographer	0.11
licensed professional counselor	0.98	language teacher	0.09	mummy	0.23	cinematographer	0.20	ally	0.95	retoucher	0.11
notary public	0.98	conductor	0.09	psychic	0.23	retoucher	0.19	cleaner	0.95	sideman	0.11

Occ. families: Arts, Design, Entertainment, Sports, & Media ■; Community & Social Service ■; Computer & Mathematical ■; Sales & Related ■

필터링 기법들이 선호하는 데이터가 분명히 존재

주제 2.

Tokenizer에도 Bias가 있다.

의미적으로 동일한 두개의 문장을

LLAMA2 tokenizer로 분절하면 무슨 일이 일어날까 ??



골프와 럭비는 모두 올림픽 게임에 복귀하기로 예정되어 있습니다.

(35 characters)



Both golf and rugby are set to return to the Olympic Games.

(59 characters)

주제 2.

Tokenizer에도 Bias가 있다.

의미적으로 동일한 두개의 문장을

LLAMA2 tokenizer로 분절하면 무슨 일이 일어날까 ??

골프와 럭비는 모두 올림픽 게임에 복귀하기로 예정되어 있습니다.

➔ [1, 29871, 237, 182, 171, 240, 151, 135, 239, 156, 131, 29871, 238, 162, 176, 31487, 31081, 29871, 31962, 238, 148, 147, 29871, 239, 155, 175, 238, 169, 191, 240, 151, 192, 29871, 237, 181, 143, 239, 161, 135, 31054, 29871, 238, 182, 184, 237, 186, 131, 30944, 30827, 30906, 29871, 239, 155, 139, 30852, 238, 147, 155, 31129, 29871, 239, 161, 139, 239, 141, 184, 31063, 30709, 29889]

69 tokens

Both golf and rugby are set to return to the Olympic Games.

➔ [1, 9134, 29416, 322, 20747, 526, 731, 304, 736, 304, 278, 19025, 12482, 29889]

14 tokens

주제 2.

Tokenizer에도 Bias가 있다.

의미적으로 동일한 두개의 문장을

GPT4에서 쓰는 tokenizer로 분절하면 무슨 일이 일어날까 ?? OpenAI Tiktoken

ས་གནས་ཀྱི་བདེ་བརྒྱུད་ནང་ལ་ གནས་ཀྱི་ཐང་གི་ མེ་བསང་མེ་ལྷུ་མ་འཁོར་འདི་ མེ་བསང་ཚོད་པའི་སྐབས་ལ་ རྩོམ་རིལ་སོང་ཡོད་པ་ཟེ་སྐད་ཀྱི་འབད་ཡོད་པ་མིན་མས། (부탄어)

→

(139 characters)

Local media reports an airport fire vehicle rolled over while responding.

→

(73 characters)

주제 2.

| Tokenizer에도 Bias가 있다.

의미적으로 동일한 두개의 문장을

GPT4에서 쓰는 tokenizer로 분절하면 무슨 일이 일어날까 ?? OpenAI Tiktoken

མ་གནས་ཀྱི་བད་བརླུང་ནང་ལ་ གནས་ཀྱི་ཐང་གི་ མེ་བསང་མེ་ལྷུ་འཁོར་འདི་ མེ་བསང་ཚོད་པའི་སྐབས་ལ་ རྩོམ་རིལ་སང་ཡོད་པ་ཟེ་སྐྱོན་ལྷན་འབད་ཡོད་པ་མིན་མས། (부탄어)

→ [63516, 99, 92988, 233, 63516, 224, 63516, 241, 63516, 99, 92988, 233, 63516, 222, 156, 122, 109, 63516, 110, 92988, 233, 63516, 244, 63516, 95, 156, 122, 94, 92988, 233, 63516, 244, 63516, 95, 156, 122, 240, 156, 122, 109, 63516, 112, 63516, 239, 92988, 233, 63516, 241, 63516, 226, 92988, 233, 63516, 96, 63516, 112, 92988, 233, 29082, 121, 224, 63516, 241, 63516, 246, 92988, 233, 63516, 224, 156, 122, 110, 63516, 112, 92988, 233, 63516, 238, 63516, 226, 92988, 233, 63516, 224, 63516, 110, 92988, 233, 29082, 121, 246, 63516, 118, 92988, 233, 63516, 244, 63516, 99, 63516, 239, 92988, 233, 63516, 246, 63516, 110, 92988, 233, 63516, 99, 156, 122, 96, 63516, 112, 63516, 246, 92988, 233, 63516, 254, 63516, 223, 63516, 120, 63516, 95, 92988, 233, 63516, 254, 63516, 239, 63516, 110, 92988, 233, 29082, 121, 246, 63516, 118, 92988, 233, 63516, 244, 63516, 99, 63516, 239, 92988, 233, 63516, 99, 156, 122, 238, 63516, 244, 63516, 99, 92988, 233, 63516, 96, 63516, 112, 92988, 233, 29082, 121, 95, 156, 122, 99, 63516, 120, 63516, 244, 92988, 233, 63516, 95, 63516, 110, 63516, 96, 92988, 233, 63516, 99, 63516, 120, 63516, 226, 92988, 233, 63516, 94, 63516, 120, 63516, 239, 63516, 242, 92988, 233, 63516, 99, 156, 122, 99, 63516, 118, 92988, 233, 63516, 99, 156, 122, 247, 63516, 241, 92988, 233, 63516, 252, 63516, 112, 92988, 233, 63516, 254, 63516, 244, 63516, 239, 92988, 233, 92988, 233, 63516, 94, 63516, 120, 63516, 239, 63516, 242, 92988, 233, 63516, 101, 63516, 110, 63516, 241, 92988, 233, 63516, 246, 63516, 99, 92988, 235]

285 tokens

Local media reports an airport fire vehicle rolled over while responding.

→ [7469, 3772, 6821, 459, 17149, 4027, 7458, 23255, 927, 1418, 30438, 13]

23.75배
차이 !!

12 tokens

주제 2. Tokenizer에도 Bias가 있다

| 어떤 문제가 있나?

예를 들어, 저 문장을 paraphrase하고 싶을 때

Paraphrase the following statement:

མ་གནས་ཀྱི་བདེ་བརྒྱད་ནང་ལ་ གནས་སྤྱོད་ཐང་གི་ མེ་བསང་མི་སྣ་འཁོར་འདི་ མེ་བསང་སྔོན་པའི་ཀླབས་ལ་ རྩོམ་པའི་སྐད་ཡིད་ལ་ རྩོམ་པའི་སྐད་ཡིད་ལ་ མིན་མས།

라고 물어보는건

Paraphrase the following statement:

Local media reports an airport fire vehicle rolled over while responding.

라고 물어보는 것보다

한 23.75배 비쌈

주제 2. Tokenizer에도 Bias가 있다

어떤 문제가 있나?

사실 실제로 해보면 23배까지는 아님 (GPT4o)

SYSTEM Paraphrase the following statements	SYSTEM Paraphrase the following statements
USER Local media reports an airport fire vehicle rolled over while responding.	USER ས་གནས་ཀྱི་བརྒྱུད་ནང་ལུ་གནས་གུ་ཐང་གི་མེ་བསང་མི་སྣུ་འཁོར་འདི་མེ་བསང་ཐོད་པའི་སྐབས་ལུ་ཐོབ་རིལ་སོང་ཡོད་པ་སྟེ་སྣེ་ལུ་འབད་ཡོད་པ་ཨིན་མས།
ASSISTANT According to local media, an airport fire truck overturned while on its way to respond.	ASSISTANT ས་གནས་ཀྱི་བརྒྱུད་ནང་ལུ་གནས་གུ་ཐང་གི་མེ་བསང་མི་སྣུ་འཁོར་འདི་མེ་བསང་ཐོད་པའི་སྐབས་ལུ་ཐོབ་རིལ་གྱིས་ཆད་འདུག་པས་ཞལ་ལག་ཐུང་པ་ཡོད།
Latency 628ms · Finish reason: stop · 46 tokens	Latency 3442ms · Finish reason: stop · 417 tokens

하지만 여전히, 똑같은 작업을 하는데 언어만 다르다는 이유로 더 많은 비용 청구
느리기도 5배쯤 더 느림

+ LM들의 Context size를 사용하는 데 제한 사항 존재
(같은 의미 단위라도, 언어에 따라 토큰 수 제한)
(RoBERTa: 512 / GPT4: 16,000)

Evaluation Measure: Tokenizer Parity

B언어 문장의 token 길이

영어에 대한 부탄어 premium은 23.75점

주제 2. Tokenizer에도 Bias가 있다

| 실험결과

unk ratio가
10%이상인 언어들

	GPT-2 RoBERTa	ChatGPT GPT-4	FlanT5
Bulgarian	5.51	2.64	—
Burmese	16.89	11.70	—
Chinese (Simplified)	3.21	1.91	—
Dzongkha	16.36	12.33	—
English	1.00	1.00	1.00
French	2.00	1.60	1.60
German	2.14	1.58	1.37
Italian	2.01	1.64	2.18
Japanese	3.00	2.30	—
Jingpho	2.65	2.35	3.41
Maori	2.45	2.35	3.28
Norwegian Bokmål	1.86	1.56	2.24
Odia	13.38	12.48	—
Pangasinan	1.66	1.57	2.18
Portuguese	1.94	1.48	2.21
Romanian	2.48	1.88	1.50
Santali	12.86	12.80	—
Shan	18.76	15.05	—
Spanish	1.99	1.55	2.23
Standard Arabic	4.40	3.04	—
Tumbuka	2.78	2.57	3.29
Vietnamese	4.54	2.45	—

번역 평가용 parallel corpus
FLORES-200 활용해서 검증

- Premium이 가장 적은 포르투갈어에서도, 영어보다 1.5배 많은 토큰이 요구됨
- 최대는 15.05배까지 요구됨 (미얀마 소수어 Shan)

Shan에서 “You”를 뜻하는 “ꠄꠤ”를 인코딩하면?

25870	247	157	224	224	25870	118	25870	116
ꠄ		ꠄ			ꠄ		ꠄ	

주제 2. Tokenizer에도 Bias가 있다

실험결과

	Arabic BERT	RoCBert (Chinese)	CamemBERT (French)	GottBERT (German)	BERT Japanese	PhoBERT (Vietnamese)
Belarusian	4.74	—	—	5.62	—	3.46
Bulgarian	4.30	—	—	4.73	—	3.09
Catalan	2.36	2.86	1.59	1.89	1.95	1.57
Chinese (Simp.)	—	1.00	—	3.95	0.82	—
Chinese (Trad.)	—	0.94	—	3.82	0.84	—
Dutch	2.52	2.92	1.68	1.73	1.98	1.58
Dzongkha	—	—	—	16.12	—	—
English	1.83	2.60	1.20	1.35	1.49	1.20
French	2.42	3.10	1.00	1.99	2.03	1.66
Friulian	2.33	2.79	1.66	1.98	1.92	1.59
German	2.63	3.12	1.85	1.00	2.04	1.67
Greek	4.93	3.00	—	6.73	—	3.73
Italian	2.58	3.10	1.63	1.93	2.04	1.60
Japanese	1.85	1.34	—	4.35	1.00	—
Jingpho	3.12	3.12	2.13	2.55	2.47	1.84
Luxembourgish	2.56	2.97	1.82	1.75	1.96	1.72
N. Lev. Arabic	1.00	—	—	6.52	—	—
Shan	—	—	—	16.88	—	—
Standard Arabic	1.00	—	—	7.03	—	—
Tagalog	2.84	3.28	2.00	2.20	2.39	1.74
Tosk Albanian	2.66	2.90	2.17	2.39	—	2.02
Tsonga	3.01	3.09	2.03	2.29	2.46	1.76
Tumbuka	3.27	3.49	2.21	2.61	—	2.00
Vietnamese	2.52	2.55	—	4.12	—	1.00
Yue Chinese	—	0.92	—	3.75	—	—

- 비영어권 언어모델의 tokenizer
영어에 대한 선호도가 존재
(영어에 대한 premium이 작음)

- Multilingual model은,
English-centric모델에 비해 작은 parity

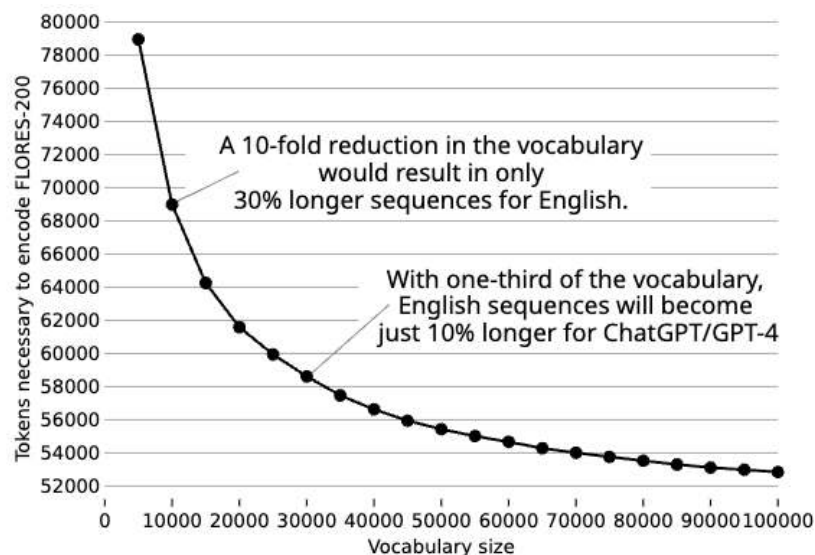
	XLM-R	NLLB	mT5	M2M100	BLOOM
Bulgarian	1.16	1.31	1.28	1.23	2.49
Central Kanuri	2.60	2.54	2.43	2.49	2.10
Chinese (Simp.)	0.97	1.11	0.92	1.05	0.95
Dzongkha	—	1.48	4.24	—	7.36
English	1.00	1.00	1.00	1.00	1.00
Indonesian	0.94	0.93	1.08	0.98	0.96
Italian	1.19	1.25	1.34	1.25	1.62
Japanese	1.11	1.01	0.90	1.20	1.81
Kabiyè	2.98	1.56	2.83	2.71	3.34
Santali	—	2.49	—	—	12.71
Shan	4.43	1.94	3.28	4.63	12.06
Std. Arabic	1.18	1.40	1.35	1.29	1.14
Std. Tibetan	—	1.44	3.68	—	6.66
Uyghur	1.41	1.40	2.57	3.00	3.67
Yue Chinese	0.93	1.05	0.95	1.03	0.93

주제 2. Tokenizer에도 Bias가 있다

제안사항 : Building a multilingually fair tokenizer from monolingual tokenizers.

- 모든 언어 데이터를 하나로 합친 Multilingual corpus에다가 tokenizer를 학습하는 것만으로는 부족하다
 - 문자 체계를 공유하는 언어들이 여럿 있기 때문에
 - "hotel"은 한국어에서만 "호텔"이고, 대부분의 영어권 언어에서는 그냥 문자 그대로 "hotel"
- 각각의 monolingual corpus로 tokenizer를 학습하고, 이후에 vocabulary를 합치는 방법을 제안
 - Dominant language에게는 기존보다 더 많은 토큰이 요구되나, 다른 언어에서 얻는 이점에 비하면 그 비용은 낮은 수준일 것

Figure 3: How much longer will English language tokenization be if we dedicate a fraction of the `cl100k_base` vocabulary to other languages? This plot shows how many tokens will be necessary to encode the English language corpus of FLORES-200 for different subsets of the `cl100k_base` vocabulary.

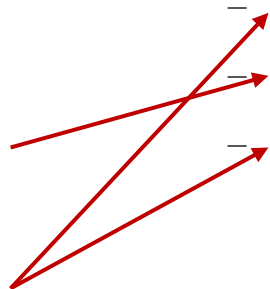


주제 2. Tokenizer에도 Bias가 있다

개인적인 실험 - 한국어LLM

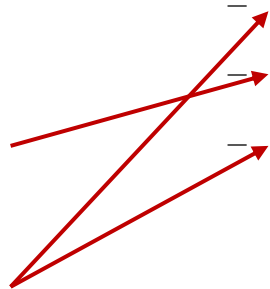
$$\text{FLORES-200} \frac{\text{영어 문자길이}}{\text{한국어 문자길이}} = 0.509$$

English-Centric LLM		Korean-Centric LLM		Multilingual LM	
- GPT4:	2.397	- KULLM2:	2.456	- BLOOM:	2.807
- LLAMA2:	3.179	- BLLOSSOM:	1.484	- XLM-R:	1.154
- LLAMA3:	1.484	- EEVE:	1.142		
- LLAMA3.1:	1.484				
- SOLAR:	2.456				



$$\text{Aihub Ko-En} \frac{\text{영어 문자길이}}{\text{한국어 문자길이}} = 0.445$$

English-Centric LLM		Korean-Centric LLM		Multilingual LM	
- GPT4:	2.133	- KULLM2:	2.162	- BLOOM:	2.482
- LLAMA2:	2.774	- BLLOSSOM:	1.316	- XLM-R:	1.019
- LLAMA3:	1.316	- EEVE:	1.004		
- LLAMA3.1:	1.316				
- SOLAR:	2.162				



| Conclusion

Bias는 어디에나 있다

찾는건 우리의 몫

감사합니다