Prior-based Noisy Text Data Filtering: Fast and Strong Alternative For Perplexity

Yeongbin Seo, Gayoung Kim, Jaehyung Kim*, Jinyoung Yeo*

Filtering noisy web text for LLM pretraining

Previous approaches

- Rule-based:
  → Efficient but limited in performance
  Ex) bad word list, minimum length
- Model-based
  → High performance but resource-intensive
  Ex) classifier, Perplexity-based

- Ours: Prior-based:
  → SOTA performance, x1000 faster than PPL

Using empirical law in linguistics

① Term frequency is a 1-dimensional representation of a word's role

Across languages, high-frequency words correspond to functional words.

Tom is a happy boy
James is the next president

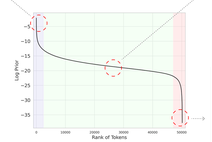
② Well-formed sentences typically exhibit a consistent range of lexical density

		content : function
Well formed ✓	Tom is a happy boy	3 : 2
	James is the next president	3 : 2
Ill formed ✗	boy boy boy boy boy	5 : 0
	is the a is the	0 : 5

Role type by **term frequency (prior)**

High prior zone (function words)

.	,	_the
\n	_to	_and
_of	_	_a
_in	-	_is
_for	_that	_with
_on	_you	âĢ
_it	_are	_as
_be	Ļ	s
_(	_|	:
_this	_have	_or

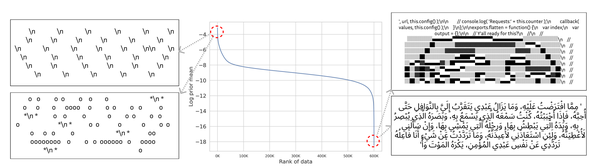


Middle prior zone (content words)

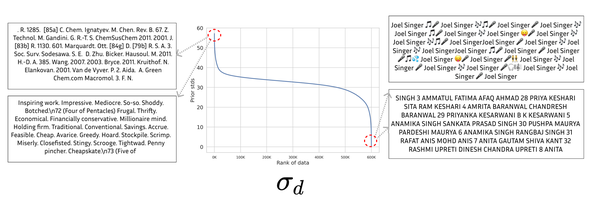
_colder	_Creation	_shortcut	_compens
423	_tackles	Phone	_outrage
_doorstep	pid	igration	onym
DN	_Kor	_inference	Authent
_slaughter	_Chester	_Fib	_pharmacies

Low prior zone (noise)

Downloadha	oreAndOnline
InstoreAndOnline	ÃĥÃĥÃĥÃĥÃĥÃĥÃĥÃĥÃĥÃĥ ÃĥÃĥÃĥÃĥÃĥÃĥÃĥÃĥÃĥÃĥ
rawdownloadcloneembedreportprint	_ãĤµãĥ¼ãĥĩãĤ£
_"$:/	_Smartstocks
=~=~	BuyableInstoreAndOnline



μ_d



σ_d

Deriving filtering criteria from linguistics

$$\mu_d = \mathbb{E}_{x_i \in d}[\log p_{prior}(x_i)]$$

$$\sigma_d = \texttt{std}_{x_i \in d}[p_{prior}(x_i)] \qquad d \in D$$

The mean μ_d and standard deviation σ_d of each token x_i are expected to cluster around a central value. Data points far from this center are likely to be ill-formed text (outliers).

Proof in a Simplified Setting

$\texttt{C}$: content words

$\texttt{F}$: function words

$$p^*(\texttt{C}) = \frac{4}{5}, \qquad p^*(\texttt{F}) = \frac{1}{5}$$

Observed data point

(Case1) CCCCF → well formed

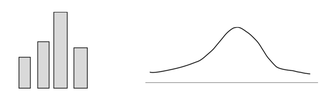
(Case2) CCFFC → ill formed

(Case3) CCCCC → ill formed

(1) In the well-formed case (p*), m and σ have fixed values.

(2) If well-formed cases dominate (at least over 10%) in the web corpora, the distribution of μ_d and σ_d will be centered around a single point.

$\mathbb{E}(\mu_d) \approx 0.68$
$\mathbb{E}(\sigma_d) \approx 0.26$



Edge case

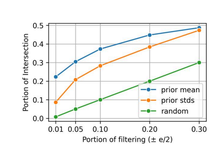
When the data has correct C:F ratio, but no meaning for the model
Ex) too difficult or too easy data

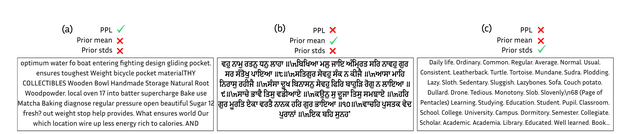
→ prior-based filter is only good for filtering "noisy" data

Relation with Perplexity-based filter

$$\log \mathrm{PPL}(\mathrm{d}) \propto \underbrace{\sum_{i}^{N} \log p_\theta(x_{<i}|x_i)}_{\pi_{likelihood}} + \underbrace{\sum_{i}^{N} \log p_\theta(x_i)}_{\pi_{prior}}$$

Weak approximation





Experiment

1. English corpus & General task

	Type	Time	Avg	W	C	L	S	R
1.5B model								
vanilla	rule-based	-	5.78	5.52	0.44	6.14	**13.22**	3.59
FastText	classifier	3.6 hrs	7.09	6.71	6.11	6.89	11.93	3.82
DSIR	n-gram	4 hrs	7.56	7.03	6.84	7.31	12.67	**3.97**
PPL-based	model-based	216 GPU hrs	8.22	**9.98**	**11.91**	7.34	7.91	3.96
Prior-based	(ours)	**0.25 hrs**	**9.20**	9.53	11.27	**10.31**	11.13	3.79
137M model								
vanilla	rule-based	-	4.96	4.96	1.81	1.47	**12.83**	3.70
FastText	classifier	3.6 hrs	5.39	5.12	4.29	1.74	12.31	3.49
DSIR	n-gram	4 hrs	5.60	**5.68**	4.93	1.97	11.60	**3.80**
PPL-based	model-based	216 GPU hours	5.26	5.47	6.53	2.90	7.84	3.58
Prior-based	(ours)	**0.25 hours**	**6.65**	5.03	**9.13**	**4.22**	11.21	3.66

Pre-trained on Dolma for 6B token duration. Evaluated across 20 benchmarks.

Avg: Avg normalized accuracy, **W:** World knowledge, **C:** Commonsense reasoning, **L:** Language understanding, **S:** Symbolic problem solving, **R:** Reading comprehension

Ours achieves better performance than PPL-based methods and is 1000 times faster.

2. Symbolic language corpus

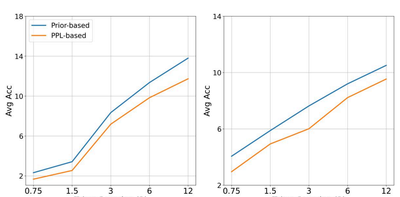
	Time	Avg	CS	Dyck	Op	Ma	GSM	SVA
1.5B model								
no-filter	-	9.51	35.75	12.30	5.71	1.15	**0.15**	**2.00**
PPL-based	224 GPU hours	11.21	37.42	20.60	7.14	**2.09**	0.00	0.00
Prior-based (ours)	**0.26 hours**	**12.03**	**38.86**	**21.30**	**9.04**	1.17	**0.15**	1.67
137M model								
no-filter	-	10.15	37.87	**16.30**	5.23	1.52	0.00	0.00
PPL-based	224 GPU hours	9.82	**40.45**	14.10	1.42	2.61	**0.07**	0.33
Prior-based (ours)	**0.26 hours**	**12.19**	40.22	16.00	**7.14**	**3.08**	0.00	**6.66**

pre-trained on Pile-github for 6B token duration, evaluated across 6 symbolic problem-solving benchmarks.

CS: BIG-bench cs algorithms, **Dyck:** BIG-bench dyck languages, **Op:** BIG-bench operators, **Ma:** BIG-bench elementary math QA, **GSM:** GSM8k, **SVA:** SVAMP

Ours demonstrates consistent advantages in symbolic language and tasks.
In addition, compared to FastText and DSIR—which require reference curation—it is much simpler to implement.

3. Large scale generalization



(a) 3B model (b) 1.5B model

Consistent superiority over PPL-based filter, on larger model size and training scale