

General overview

Corpus	Date	Language
hplt-v3-ind_Latn	9/18/2025	Indonesian (id)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
176,107,709	3,542,041,702	2,333,497,780 (65.88 %)	101B	607,150,200,712	567.75 GB

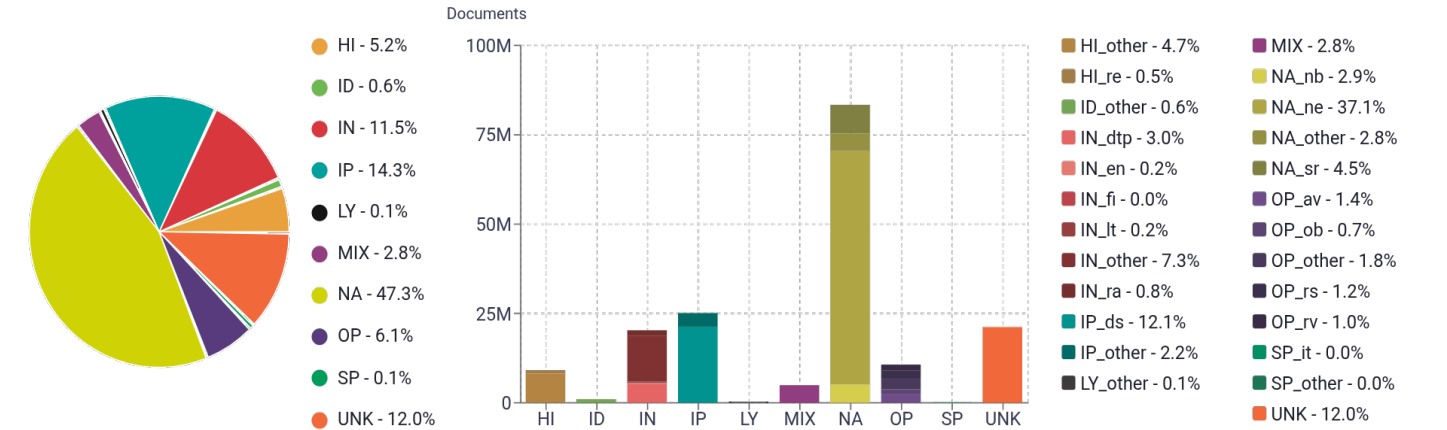
Top 10 domains

Domain	Docs	% of total
blogspot.com	5.2M	2.93%
tribunnews.com	4.3M	2.45%
wordpress.com	3.5M	1.99%
kompas.com	2M	1.13%
tempo.co	769K	0.44%
detik.com	757K	0.43%
blogspot.co.id	693K	0.39%
republika.co.id	647K	0.37%
okezone.com	637K	0.36%
antaranews.com	534K	0.30%

Top 10 TLDs

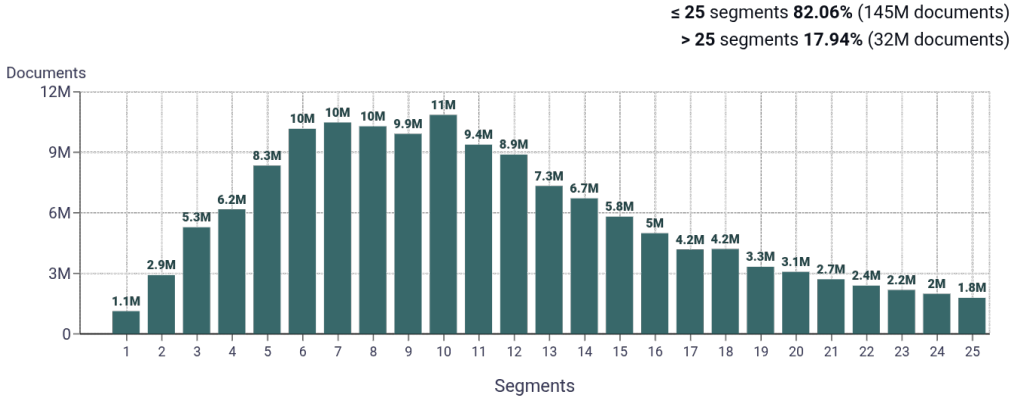
Domain	Docs	% of total
com	108M	61.41%
id	13M	7.28%
co.id	11M	6.49%
net	6.6M	3.76%
org	5.4M	3.08%
co	4.4M	2.49%
ac.id	3.7M	2.12%
go.id	3M	1.69%
info	2.5M	1.39%
xyz	1.2M	0.70%

Register labels

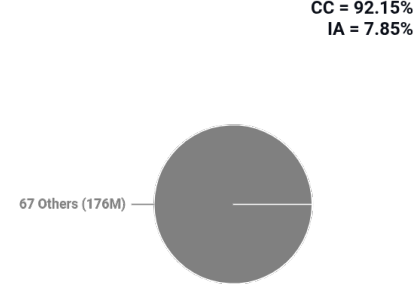


MT:6.3% | 11M Documents

Documents size (in segments) ⓘ

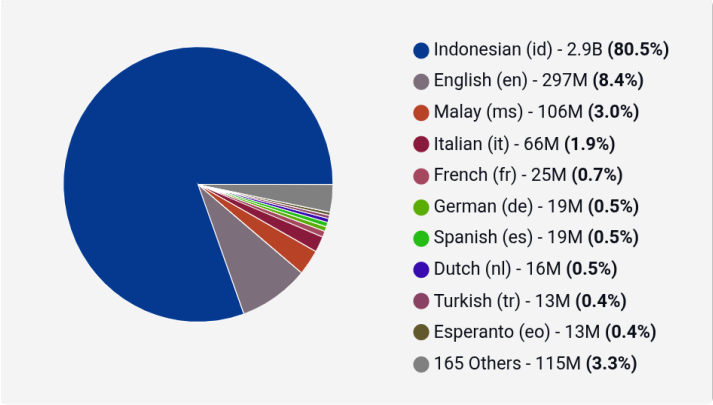


Document collections

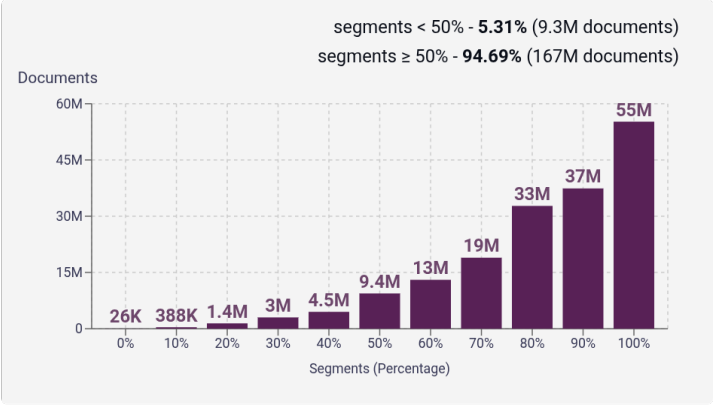


Language Distribution

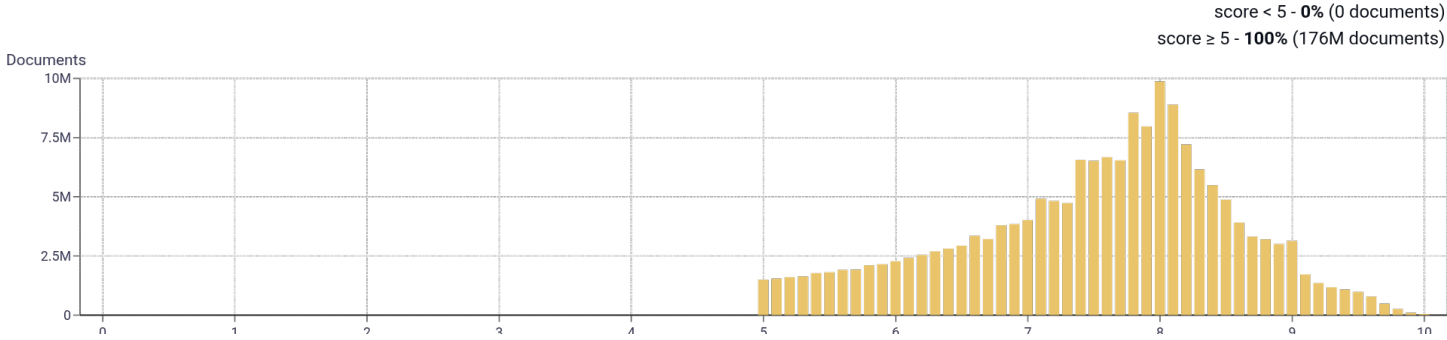
Number of segments in the Indonesian (id) corpus



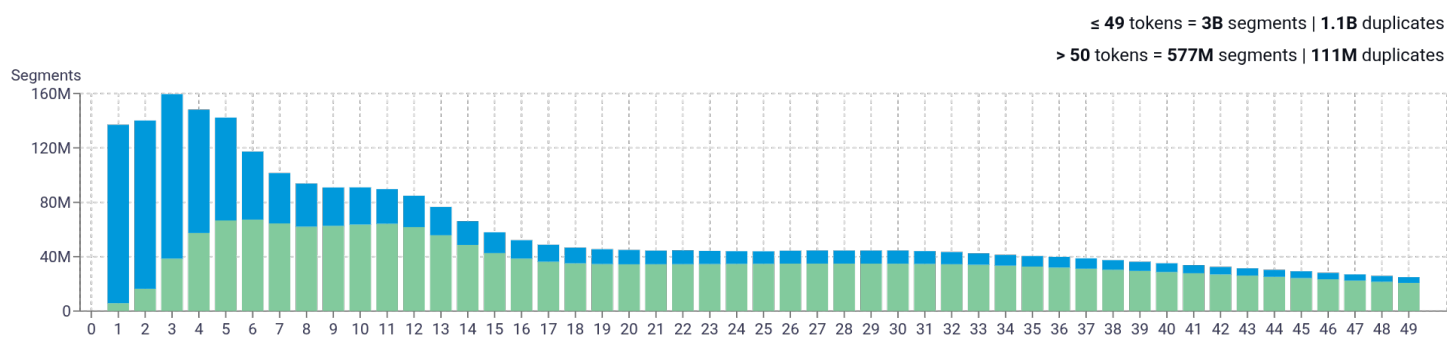
Percentage of segments in Indonesian (id) inside documents



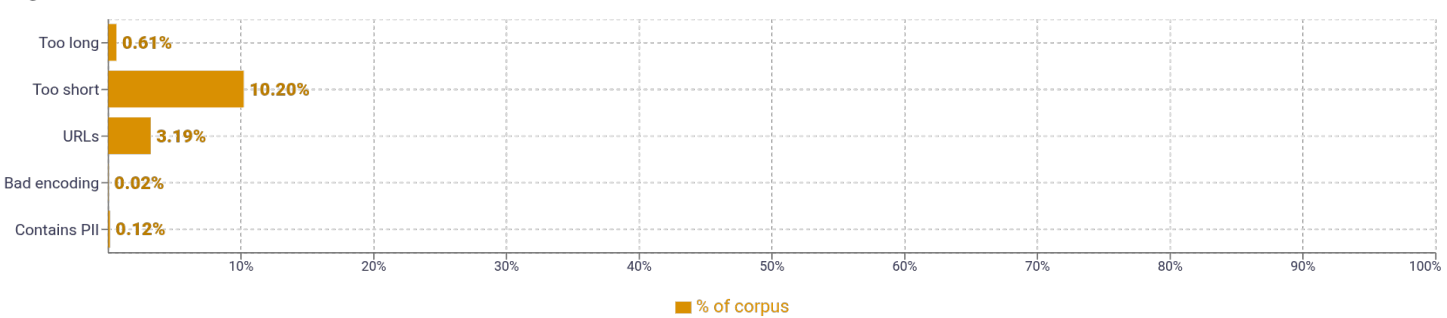
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>nya 1,322,049,037</div> <div>online 202,653,703</div> <div>indonesia 188,723,623</div> <div>memiliki 185,980,381</div> <div>orang 182,998,563</div>	
2	<div>slot online 32,926,116</div> <div>judi online 23,905,218</div> <div>situs judi 19,223,294</div> <div>poker online 16,173,000</div> <div>permainan slot 15,892,086</div>	
3	<div>salah satu nya 14,418,163</div> <div>kode promo kasino 7,920,347</div> <div>judi slot online 6,976,440</div> <div>situs judi slot 5,604,832</div> <div>gratis tanpa deposit 5,357,017</div>	
4	<div>situs judi slot online 2,909,619</div> <div>kode promo kasino online 2,727,938</div> <div>gratis kode promo kasino 2,304,697</div> <div>bonus gratis tanpa deposit 2,186,372</div> <div>lengkapi data diri mu 1,437,972</div>	
5	<div>mu untuk ikutan program #jernihberkomentar 964,840</div> <div>data diri mu untuk ikutan 964,793</div> <div>gratis bonus gratis tanpa deposit 854,494</div> <div>gratis kode promo kasino online 847,890</div> <div>kali ini kita akan membahas 739,757</div>	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , etc.) replaced by newlines.

Volumes - Tokens

Tokenized with <https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md>

Type-Token Ratio

Lexical variety computed as $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, after removing punctuation (<https://www.sltinfo.com/wp-content/uploads/2014/01/type-token-ratio.pdf>).

Document size (in segments)

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , etc.) replaced by newlines.

Language distribution

Language identified with FastSpell (<https://github.com/mbanon/fastspell>).

Distribution of segments by fluency score

Obtained with Monocleaner (<https://github.com/bitextor/monocleaner>).

Distribution of documents by average fluency score

Obtained with Monocleaner (<https://github.com/bitextor/monocleaner>).

Distribution of documents by document score

Obtained with Web Docs Scorer (<https://github.com/pablop16n/web-docs-scorer/>).

Segment length distribution by token

Tokenized with <https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md>

Segment noise distribution

Obtained with Bicleaner Hardrules (<https://github.com/bitextor/bicleaner-hardrules/>).

Frequent n-grams

Tokenized with <https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md>, after removing n-grams starting or ending in a stopword. Stopwords from <https://github.com/hplt-project/data-analytics-tool/blob/main/scripts/resources/README.txt>

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	How-to or instructions	HI	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	Informational persuasion	IP	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	Opinion	OP
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	Informational description	IN	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				