HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-min_Latn	9/18/2025	Minangkabau (min)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
29,395	596,626	437,268 (73.29 %)	16M	84,541,218	81 MB

Top 10 domains

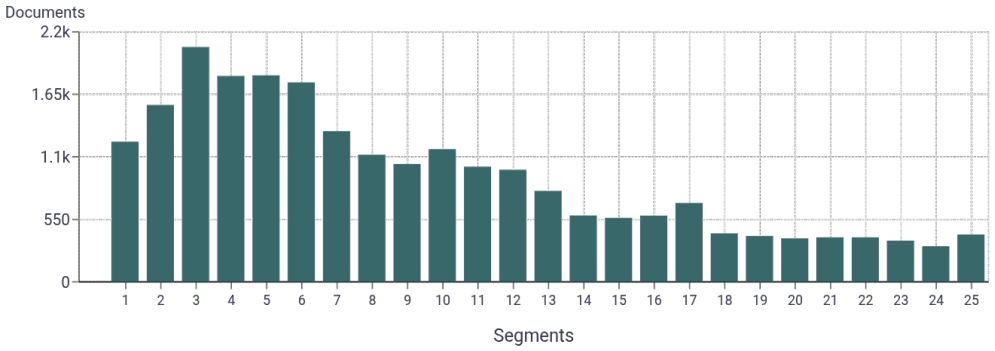
Domain	Docs	% of total
wikipedia.org	4.6K	15.80%
wordpress.com	1K	3.46%
blogspot.com	755	2.57%
minangkabaunews...	599	2.04%
minangsatu.com	556	1.89%
indonesiachord.com	489	1.66%
uinib.ac.id	438	1.49%
petalokasi.org	420	1.43%
pp.ua	341	1.16%
langgam.id	295	1.00%

Top 10 TLDs

Domain	Docs	% of total
com	10K	34.80%
org	5.9K	20.10%
ac.id	4.6K	15.65%
go.id	2.2K	7.58%
id	1.3K	4.35%
co.id	789	2.68%
net	640	2.18%
info	441	1.50%
ua	341	1.16%
is	260	0.88%

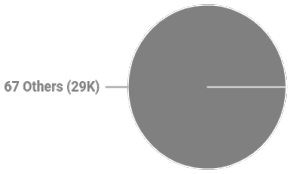
Documents size (in segments)

≤ 25 segments **78.92%** (23K documents)
> 25 segments **21.08%** (6.2K documents)



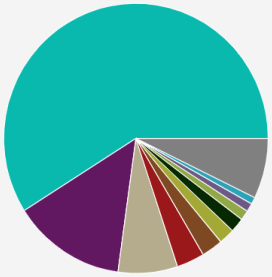
Document collections

CC = 92.75%
IA = 7.25%



Language Distribution

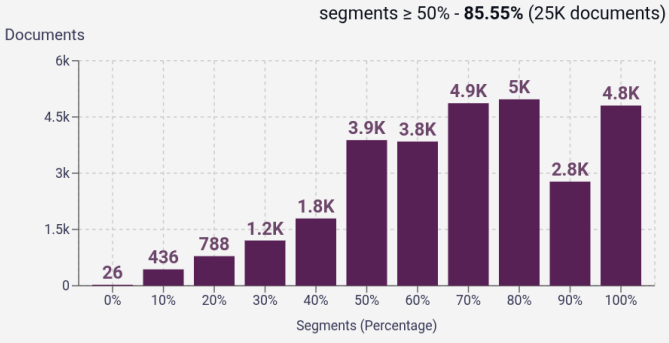
Number of segments in the Minangkabau (min) corpus



- Indonesian (id) - 352K **(59.0%)**
- English (en) - 83K **(13.8%)**
- Malay (ms) - 43K **(7.3%)**
- Minangkabau (min) - 20K **(3.4%)**
- Italian (it) - 16K **(2.6%)**
- Filipino (tl) - 12K **(1.9%)**
- Latin (la) - 10K **(1.7%)**
- Spanish (es) - 7.1K **(1.2%)**
- Javanese (jv) - 6.1K **(1.0%)**
- French (fr) - 4.9K **(0.8%)**
- 144 Others - 43K **(7.2%)**

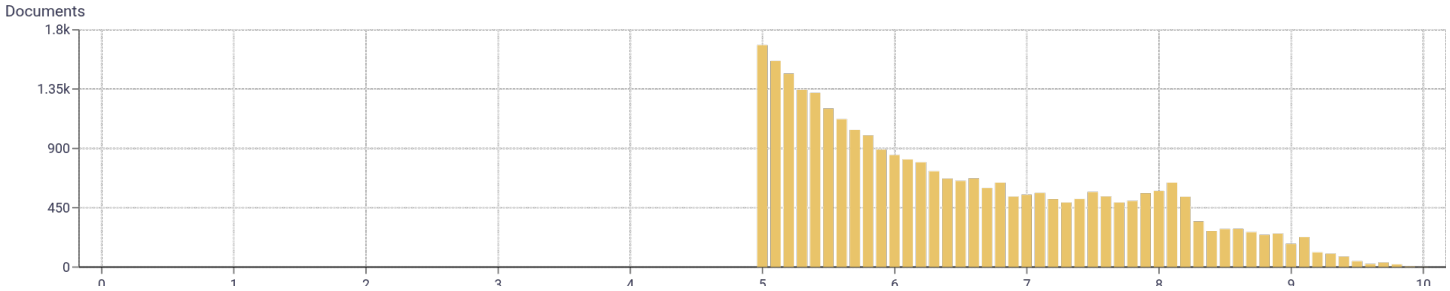
Percentage of segments in Minangkabau (min) inside documents

segments < 50% - **14.45%** (4.2K documents)
segments ≥ 50% - **85.55%** (25K documents)



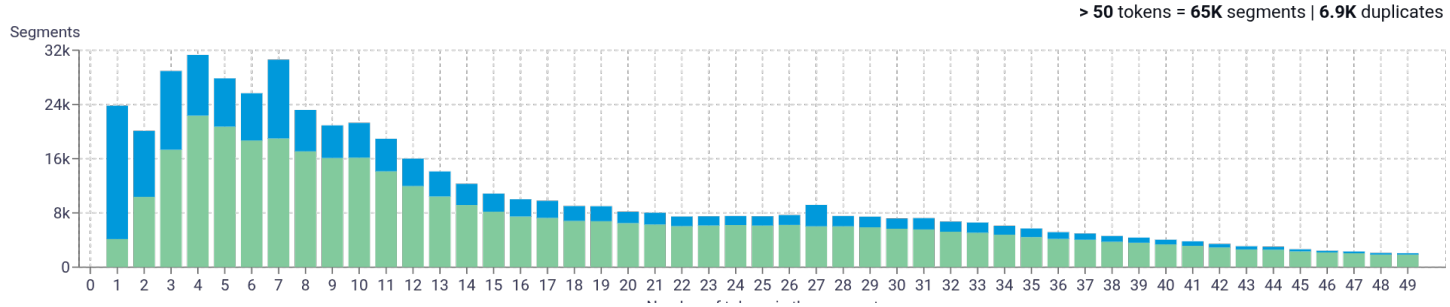
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (29K documents)

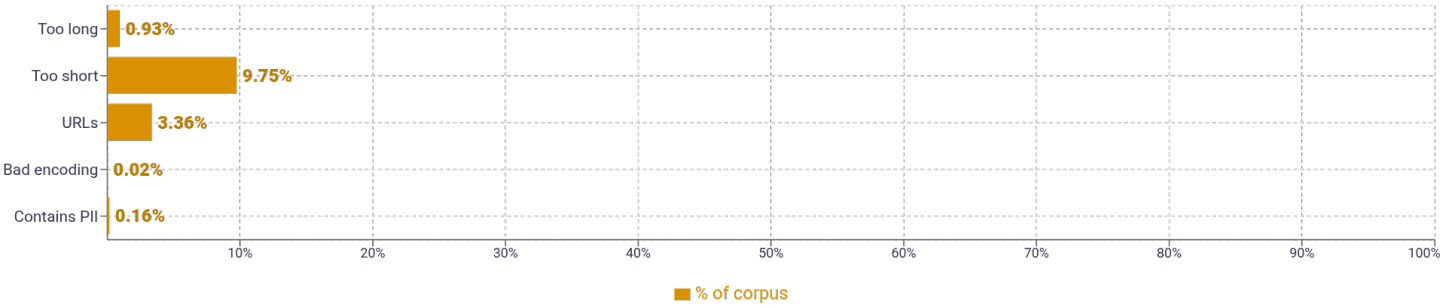


Segment length distribution by token

≤ 49 tokens = **531K** segments | **153K** duplicates
> 50 tokens = **65K** segments | **6.9K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	di | 211,302 · nan | 183,727 · jo | 93,144 · indonesia | 92,773 · padang | 77,023
2	sumatera barat | 23,788 · skripsi thesis | 19,672 · imam bonjol | 16,302 · bonjol padang | 14,581 · islam negeri | 13,864
3	imam bonjol padang | 14,545 · uin imam bonjol | 10,637 · universitas islam negeri | 10,485 · universitas negeri padang | 7,909 · pgri sumatera barat | 5,544
4	uin imam bonjol padang | 10,219 · stkip pgri sumatera barat | 5,494 · islam negeri imam bonjol | 3,905 · negeri imam bonjol padang | 3,681 · institut seni indonesia padangpanjang | 3,681
5	islam negeri imam bonjol padang | 3,669 · universitas islam negeri imam bonjol | 3,157 · universitas islam negeri sumatera utara | 2,514 · islam negeri sultan syarif kasim | 1,989 · negeri sultan syarif kasim riau | 1,988

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, 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 *number or types (uniques)/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>, <ul>, <ol>, 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				