HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
min_Latn.jsonl.tsv	12/4/2024	Minangkabau (min)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
25,037	600,798	302,936 (50.42 %)	14M	74,198,336	71.07 MB

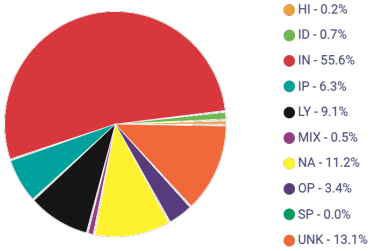
Top 10 domains

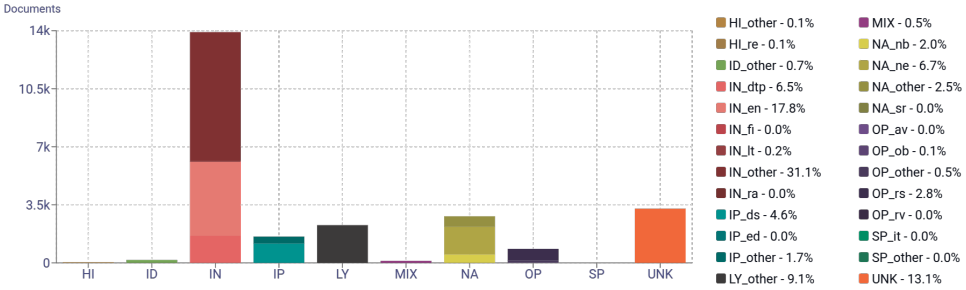
Domain	Docs	% of total
wikipedia.org	4.8K	19.25%
petalokasi.org	2.9K	11.48%
wordpress.com	1.5K	5.89%
blogspot.com	1.3K	5.12%
bible.is	639	2.55%
textmap.asia	631	2.52%
kodeposindo.xyz	613	2.45%
adatnusantara.w...	423	1.69%
chordtela.com	273	1.09%
uin-suska.ac.id	236	0.94%

Top 10 TLDs

Domain	Docs	% of total
org	8.1K	32.21%
com	7.5K	29.79%
ac.id	2.8K	11.18%
go.id	919	3.67%
asia	840	3.36%
xyz	652	2.60%
is	639	2.55%
id	585	2.34%
net	555	2.22%
web.id	499	1.99%

Register labels

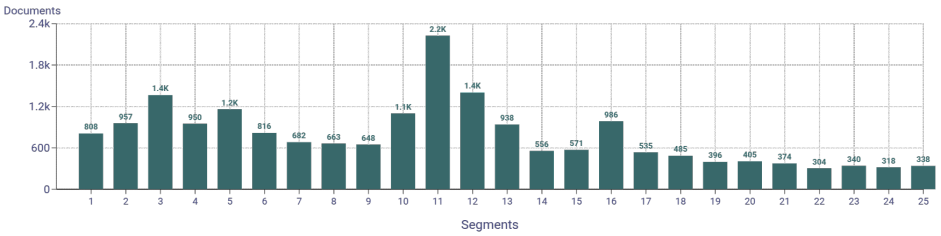




MT:2.6% | 641 Documents

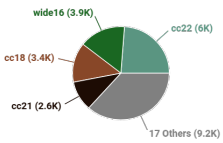
Documents size (in segments)

<= 25 segments **77.17%** (19K documents)
> 25 segments **22.83%** (5.7K documents)



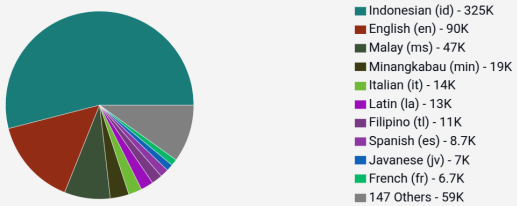
Documents by collection

CC = 64.14%
IA = 35.86%

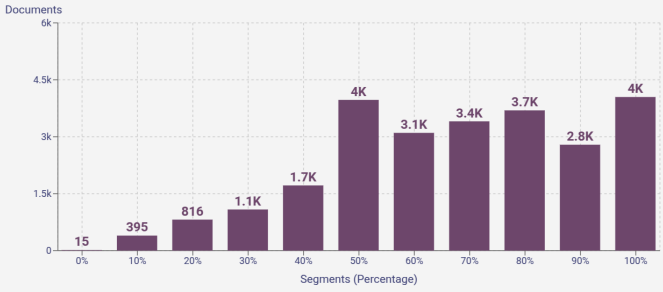


Language Distribution

Number of segments in the Minangkabau (min) corpus

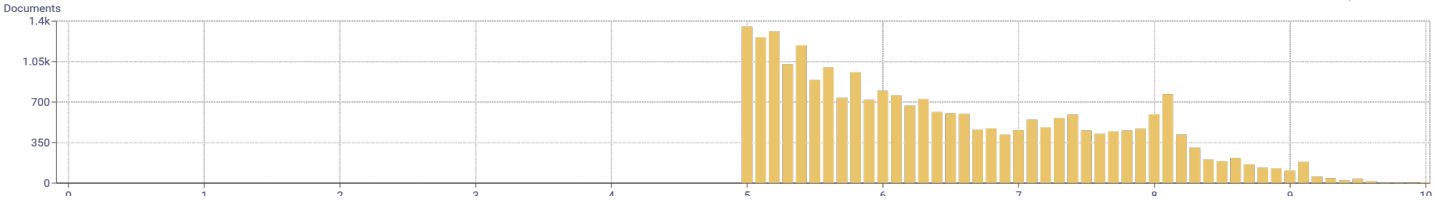


Percentage of segments in Minangkabau (min) inside documents



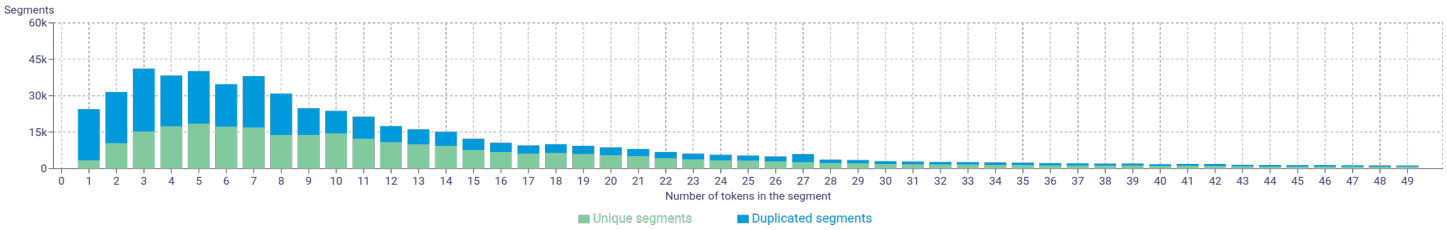
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (25K documents)

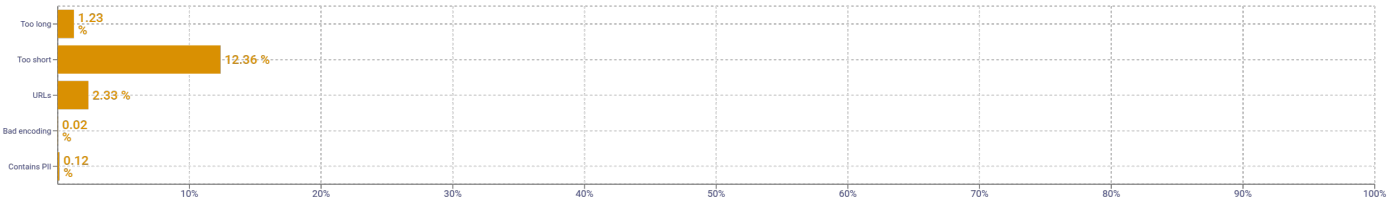


Segment length distribution by token

≤ 49 tokens = **273K** segments | **274K** duplicates
> 50 tokens = **54K** segments | **24K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	nan | 201712 · di | 169108 · indonesia | 113575 · jo | 103084 · ka | 72589
2	lagu daerah | 44250 · sumatera barat | 18121 · program studi | 16322 · suntiang sumber | 13322 · urang nan | 12965
3	kode pos kelurahan | 5428 · universitas islam negeri | 5387 · nomor kode pos | 5379 · informasi kode pos | 5367 · dj lagu daerah | 4230
4	informasi kode pos kelurahan | 5367 · hotel taj mansingh delhi | 2920 · provinsi sumatera utara nomor | 2909 · utara nomor kode pos | 2891 · sumatera utara nomor kode | 2891
5	sumatera utara nomor kode pos | 2891 · provinsi sumatera utara nomor kode | 2891 · universitas islam negeri sultan syarif | 2041 · islam negeri sultan syarif kasim | 2035 · negeri sultan syarif kasim riau | 1919

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av