HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-sun_Latn	9/18/2025	Sundanese (su)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
185,376	4,109,703	3,186,723 (77.54 %)	110M	634,166,007	612.92 MB

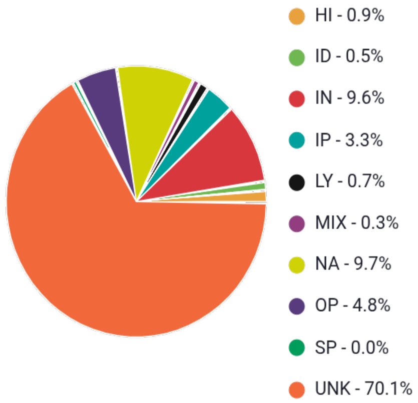
Top 10 domains

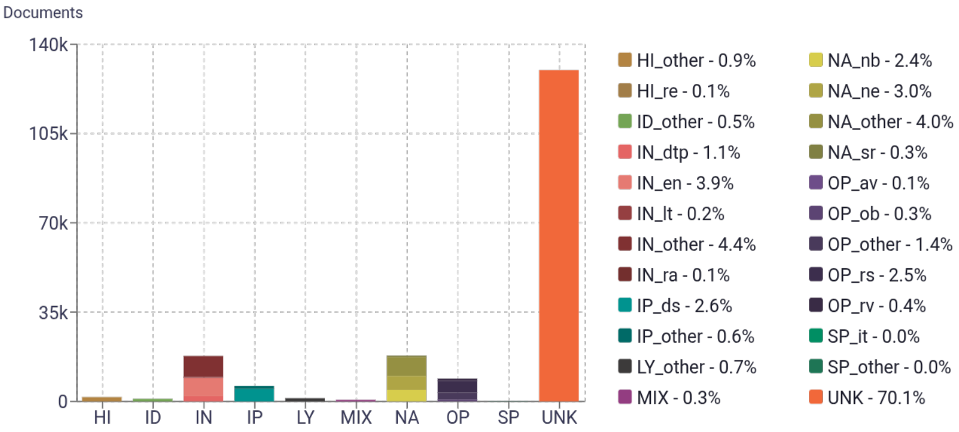
Domain	Docs	% of total
wikipedia.org	6.8K	3.69%
wordpress.com	4.9K	2.64%
martech.zone	4.8K	2.61%
sundanet.com	3.9K	2.09%
blogspot.com	3.8K	2.04%
sudanesebonus.eu	3.3K	1.76%
eturbonews.com	3K	1.62%
actualidadiphon...	2.8K	1.50%
ihorror.com	2K	1.07%
sundanews.com	1.5K	0.83%

Top 10 TLDs

Domain	Docs	% of total
com	126K	68.01%
org	13K	7.25%
icu	8.6K	4.66%
net	4.9K	2.62%
zone	4.8K	2.61%
eu	3.8K	2.03%
info	1.2K	0.64%
go.id	1.1K	0.60%
fr	994	0.54%
es	944	0.51%

Register labels

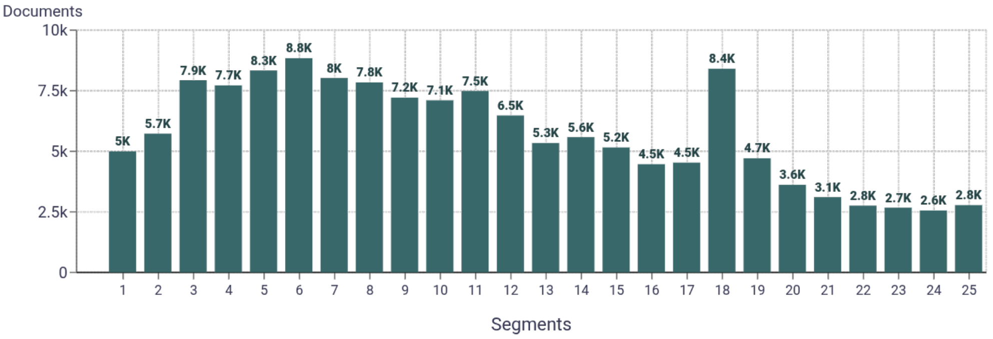




MT:67.4% | 125K Documents

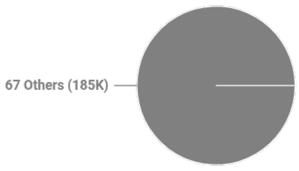
Documents size (in segments)

≤ **25** segments **77.36%** (143K documents)
> **25** segments **22.64%** (42K documents)



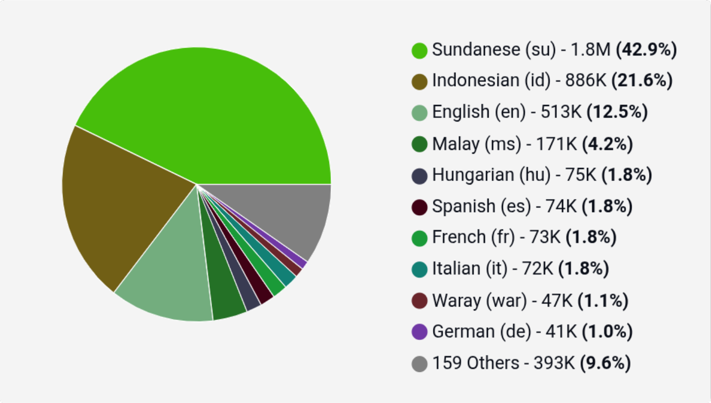
Document collections

CC = 94.30%
IA = 5.70%



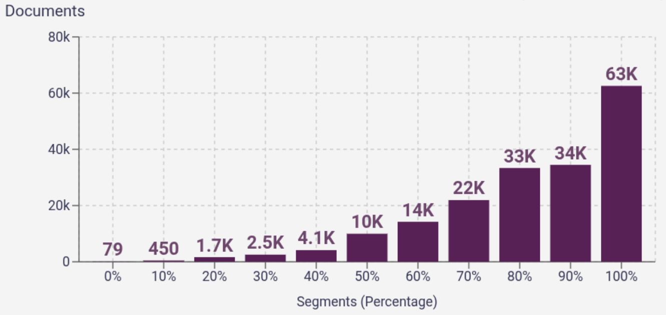
Language Distribution

Number of segments in the Sundanese (su) corpus



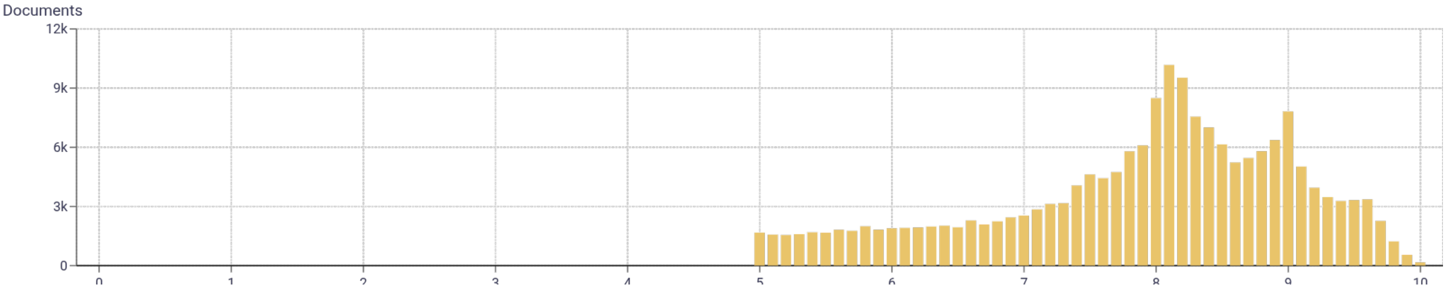
Percentage of segments in Sundanese (su) inside documents

segments < 50% - **4.77%** (8.8K documents)
segments ≥ 50% - **95.23%** (177K documents)



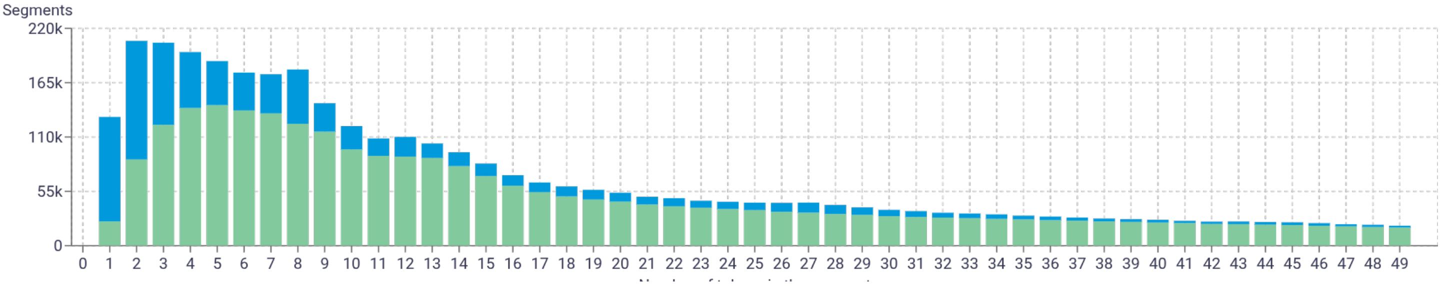
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (185K documents)

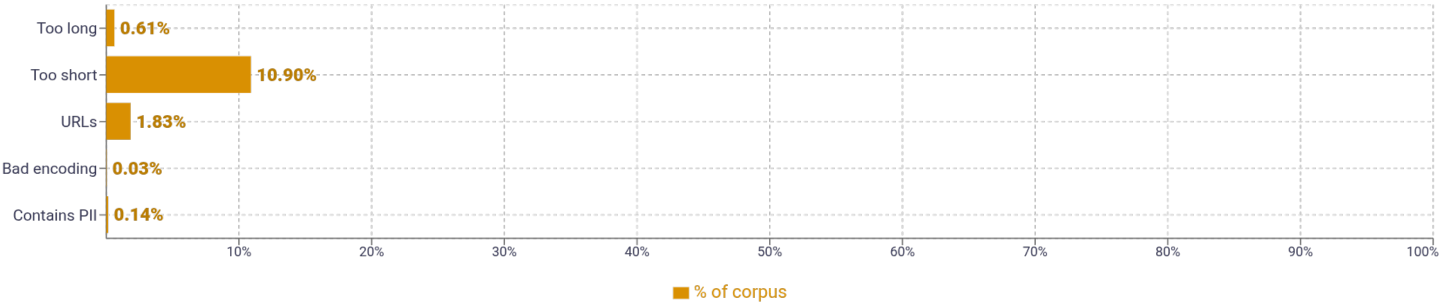


Segment length distribution by token

≤ 49 tokens = **3.5M** segments | **865K** duplicates
> 50 tokens = **627K** segments | **61K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	nu | 626,064 online | 544,007 urang | 529,486 kana | 347,854 kuring | 346,007
2	kodeu promo | 120,911 kaulinan online | 110,839 kasino kodeu | 104,468 mesin slot | 95,587 bebas kaulinan | 95,063
3	kasino kodeu promo | 99,458 online kasino kodeu | 57,986 bebas kaulinan online | 47,053 kaulinan online gratis | 27,186 kaulinan online bebas | 21,645
4	online kasino kodeu promo | 55,083 kasino kodeu promo bebas | 19,920 kodeu promo bebas kaulinan | 10,469 bebas kaulinan online on | 8,788 bebas mesin slot online | 8,745
5	online kasino kodeu promo bebas | 14,121 kasino kodeu promo bebas kaulinan | 10,469 bebas mesin slot online kalawan | 7,571 carana maén bebas kaulinan online | 7,442 mesin slot online kalawan rounds | 5,197

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				