HPLT Analytics report

HPLTAnalytics

General overview

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

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

- HI - 0.9%
- ID - 0.5%
- IN - 9.6%
- IP - 3.3%
- LY - 0.7%
- MIX - 0.3%
- NA - 9.7%
- OP - 4.8%
- SP - 0.0%
- UNK - 70.1%

- HI_other - 0.9%
- HI_re - 0.1%
- ID_other - 0.5%
- IN_dtp - 1.1%
- IN_en - 3.9%
- IN_lt - 0.2%
- IN_other - 4.4%
- IN_ra - 0.1%
- IP_ds - 2.6%
- IP_other - 0.6%
- LY_other - 0.7%
- MIX - 0.3%
- NA_nb - 2.4%
- NA_ne - 3.0%
- NA_other - 4.0%
- NA_sr - 0.3%
- OP_av - 0.1%
- OP_ob - 0.3%
- OP_other - 1.4%
- OP_rs - 2.5%
- OP_rv - 0.4%
- SP_it - 0.0%
- SP_other - 0.0%
- UNK - 70.1%
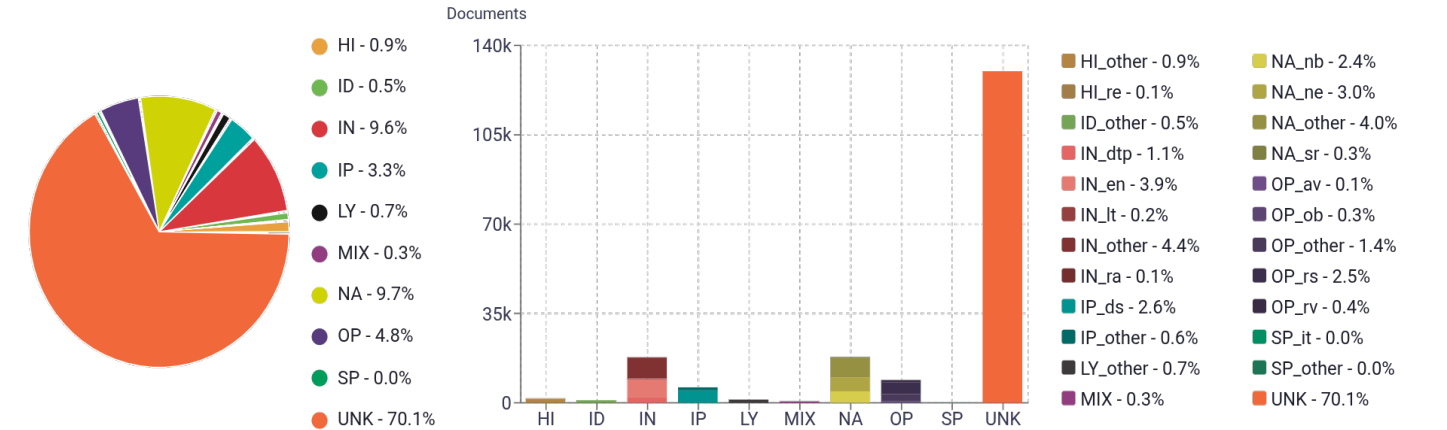
🤖 **MT**:67.4% | 125K Documents

Documents size (in segments) ⓘ

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


Document collections

CC = 94.30%
IA = 5.70%

67 Others (185K)
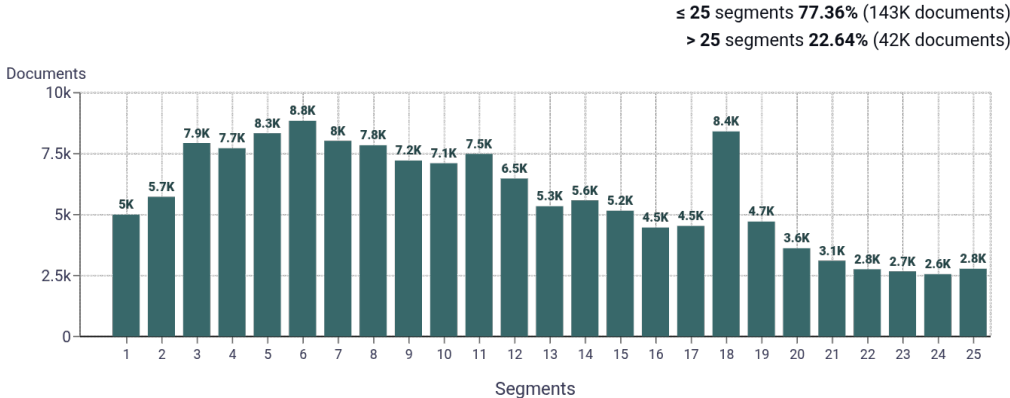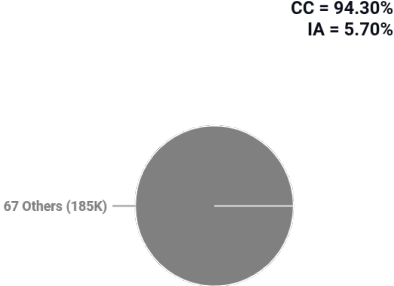
Language Distribution

Number of segments in the Sundanese corpus

- Sundanese - 1.8M **(42.9%)**
- Indonesian - 886K **(21.6%)**
- English - 513K **(12.5%)**
- Malay - 171K **(4.2%)**
- Hungarian - 75K **(1.8%)**
- Spanish - 74K **(1.8%)**
- French - 73K **(1.8%)**
- Italian - 72K **(1.8%)**
- Waray - 47K **(1.1%)**
- German - 41K **(1.0%)**
- 159 Others - 393K **(9.6%)**
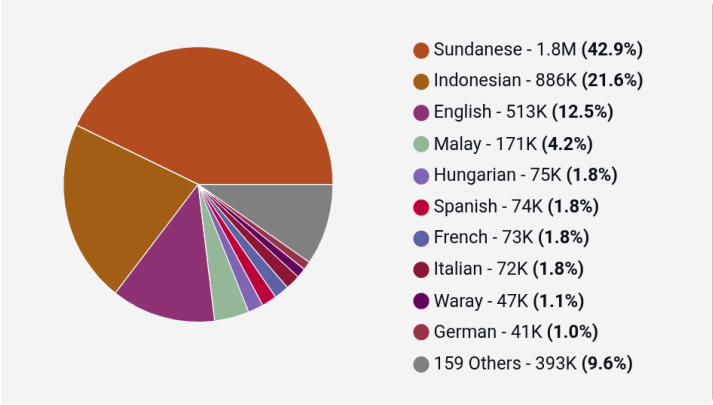
Percentage of segments in Sundanese inside documents
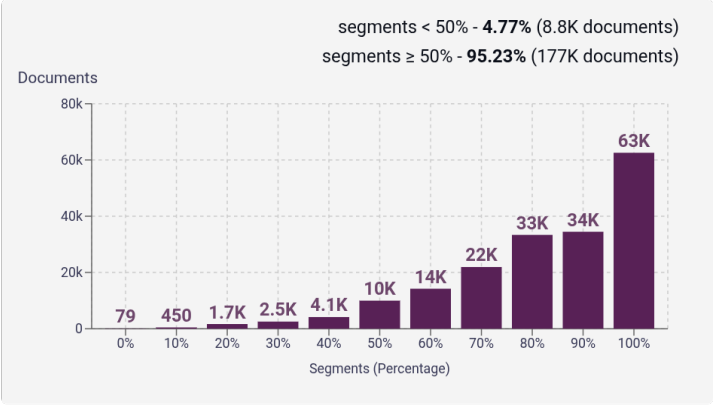
segments < 50% - **4.77%** (8.8K documents)
segments ≥ 50% - **95.23%** (177K documents)

Documents

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
79	450	1.7K	2.5K	4.1K	10K	14K	22K	33K	34K	63K

Segments (Percentage)

Distribution of documents by document score

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

Documents

Segment length distribution by token

≤ 49 tokens = **3.5M** segments | **865K** duplicates

> 50 tokens = **627K** segments | **61K** duplicates

Segments

Number of tokens in the segment

Segment noise distribution

Too long	0.61%
Too short	10.90%
URLs	1.83%
Bad encoding	0.03%
Contains PII	0.14%

■ % of corpus
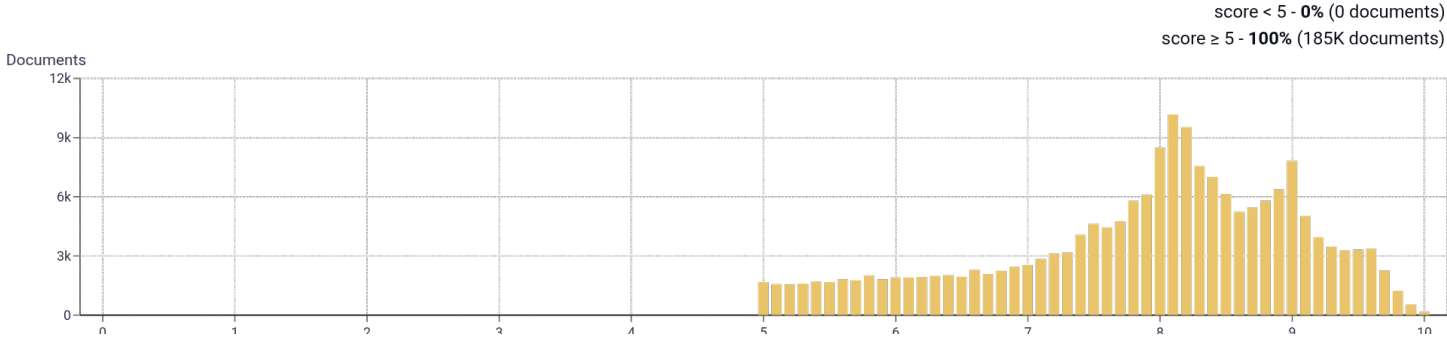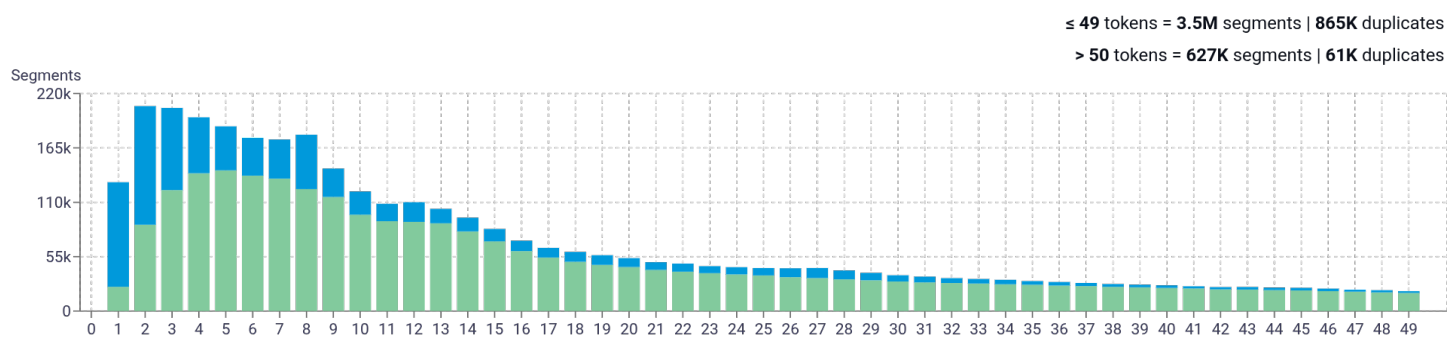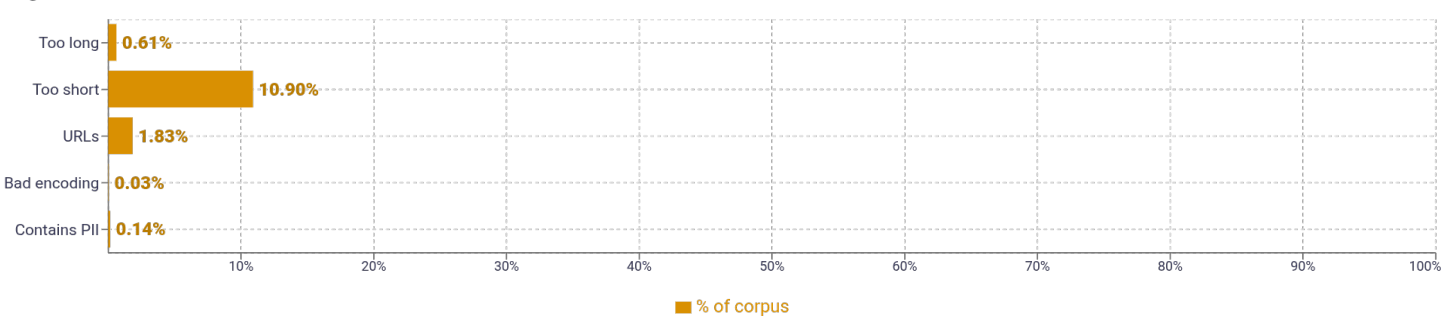
Frequent n-grams

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				