HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-zsm_Latn	9/18/2025	Malay (ms)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
17,365,290	502,955,646	295,458,173 (58.74 %)	13B	70,741,014,024	66.18 GB

Top 10 domains

Domain	Docs	% of total
blogspot.com	5.1M	29.55%
wordpress.com	311K	1.79%
blogspot.my	274K	1.58%
sinarharian.com.my	162K	0.93%
astroawani.com	144K	0.83%
mstar.com.my	118K	0.68%
utusan.com.my	113K	0.65%
hotels.com	97K	0.56%
wikipedia.org	95K	0.55%
hmetro.com.my	93K	0.53%

Top 10 TLDs

Domain	Docs	% of total
com	12M	66.96%
my	1.6M	9.26%
com.my	1.3M	7.74%
net	559K	3.22%
org	350K	2.01%
gov.my	223K	1.28%
sg	143K	0.82%
info	141K	0.81%
edu.my	139K	0.80%
com.bn	93K	0.53%

Register labels

Pie chart legend:
- HI - 3.6%
- ID - 0.9%
- IN - 8.9%
- IP - 9.7%
- LY - 0.1%
- MIX - 3.1%
- NA - 52.9%
- OP - 9.9%
- SP - 0.1%
- UNK - 10.7%

Bar chart legend:
- HI_other - 1.9%
- HI_re - 1.7%
- ID_other - 0.9%
- IN_dtp - 2.5%
- IN_en - 0.7%
- IN_fi - 0.0%
- IN_lt - 0.2%
- IN_other - 5.3%
- IN_ra - 0.2%
- IP_ds - 8.1%
- IP_other - 1.7%
- LY_other - 0.1%
- MIX - 3.1%
- NA_nb - 18.8%
- NA_ne - 27.7%
- NA_other - 4.0%
- NA_sr - 2.3%
- OP_av - 0.7%
- OP_ob - 4.2%
- OP_other - 3.0%
- OP_rs - 1.2%
- OP_rv - 0.8%
- SP_it - 0.1%
- SP_other - 0.1%
- UNK - 10.7%
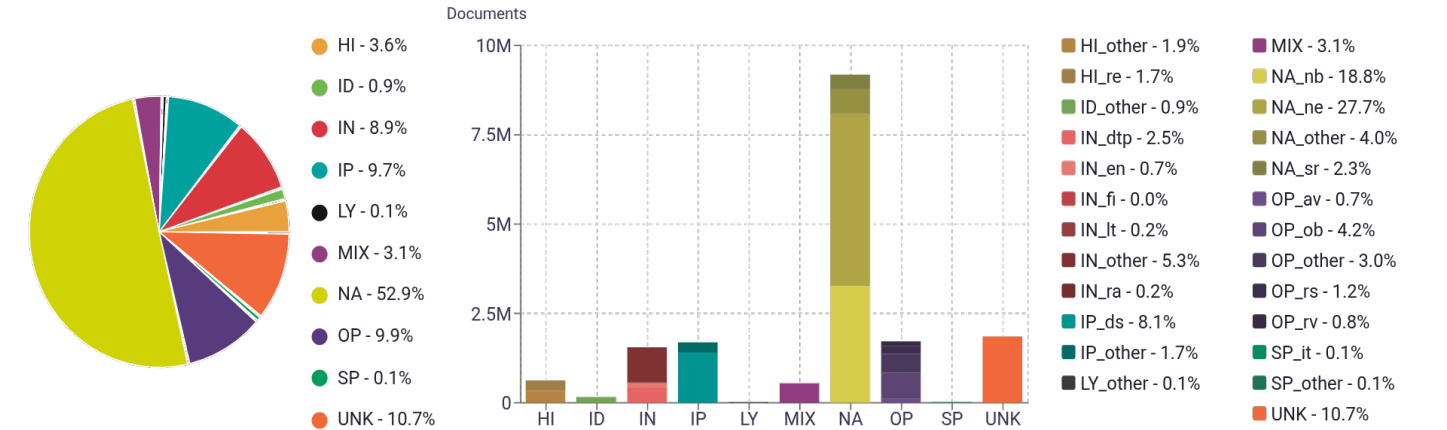
🤖 **MT**:6.7% | 1.2M Documents

Documents size (in segments) ⓘ

≤ 25 segments **73.53%** (13M documents)
> 25 segments **26.47%** (4.6M documents)

Bar chart values by segment:
1: 167K, 2: 295K, 3: 466K, 4: 555K, 5: 721K, 6: 780K, 7: 842K, 8: 863K, 9: 854K, 10: 896K, 11: 790K, 12: 705K, 13: 623K, 14: 564K, 15: 500K, 16: 453K, 17: 413K, 18: 374K, 19: 339K, 20: 322K, 21: 294K, 22: 271K, 23: 244K, 24: 225K, 25: 212K

Document collections

CC = 86.26%
IA = 13.74%

CC-MAIN-20...
66 Others (15M)
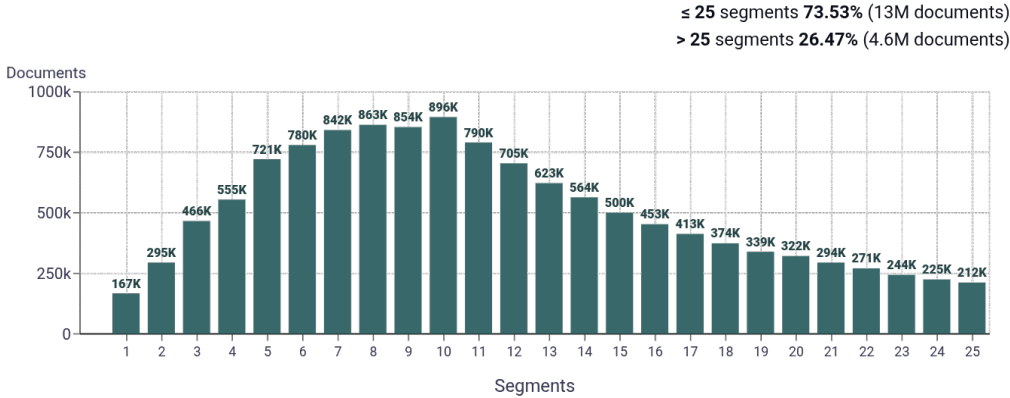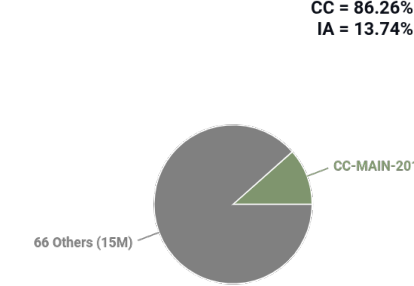
Language Distribution

Number of segments in the Malay (ms) corpus

- Malay (ms) - 333M **(66.2%)**
- English (en) - 66M **(13.1%)**
- Indonesian (id) - 57M **(11.2%)**
- Italian (it) - 8.9M **(1.8%)**
- French (fr) - 4.9M **(1.0%)**
- Spanish (es) - 3.3M **(0.7%)**
- German (de) - 3.2M **(0.6%)**
- Dutch (nl) - 2M **(0.4%)**
- Turkish (tr) - 1.6M **(0.3%)**
- Hungarian (hu) - 1.6M **(0.3%)**
- 165 Others - 22M **(4.4%)**
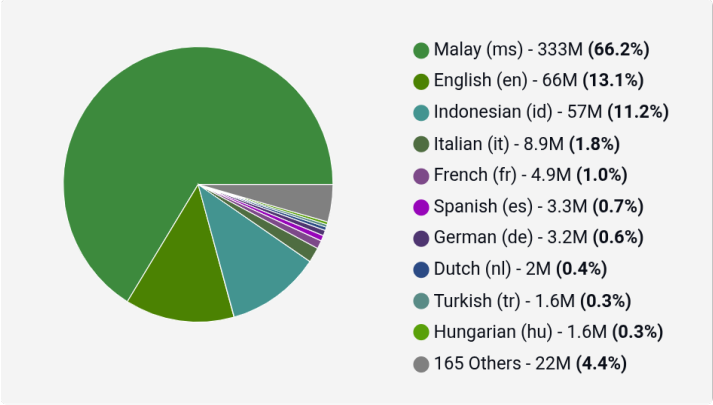
Percentage of segments in Malay (ms) inside documents
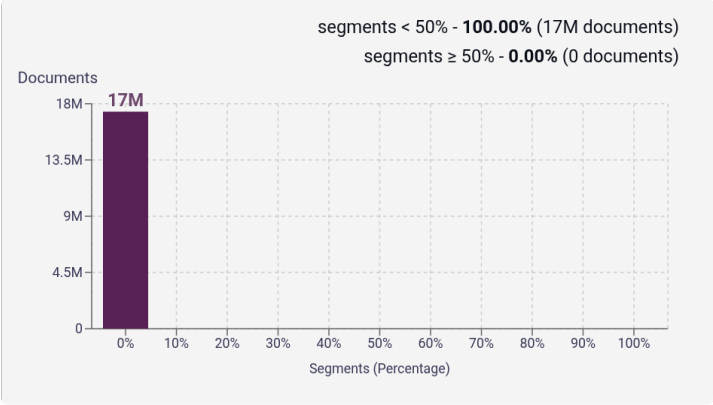
segments < 50% - **100.00%** (17M documents)
segments ≥ 50% - **0.00%** (0 documents)

Documents

17M

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (17M documents)

Documents

Segment length distribution by token

≤ **49** tokens = **444M** segments | **192M** duplicates
> **50** tokens = **58M** segments | **16M** duplicates

Segments

Segment noise distribution

- Too long — **0.40%**
- Too short — **11.94%**
- URLs — **4.43%**
- Bad encoding — **0.01%**
- Contains PII — **0.19%**

■ % of corpus
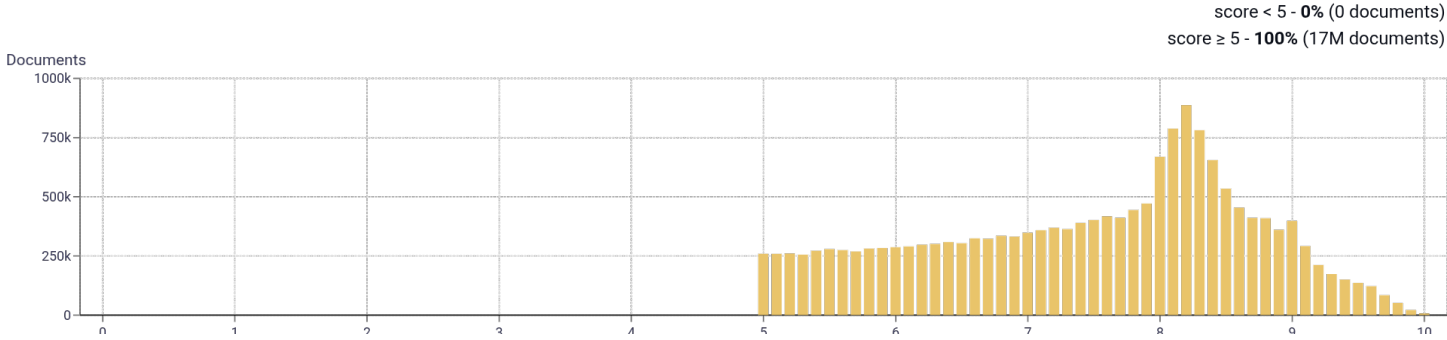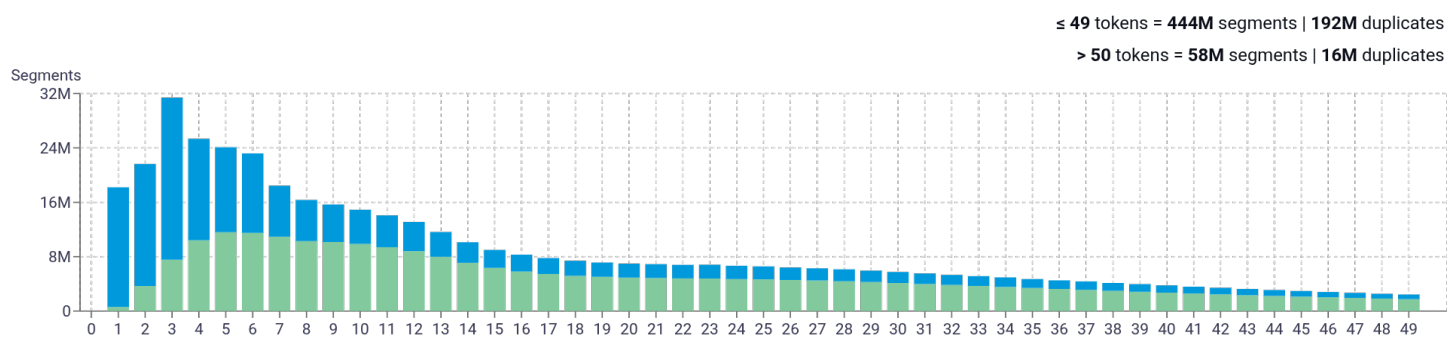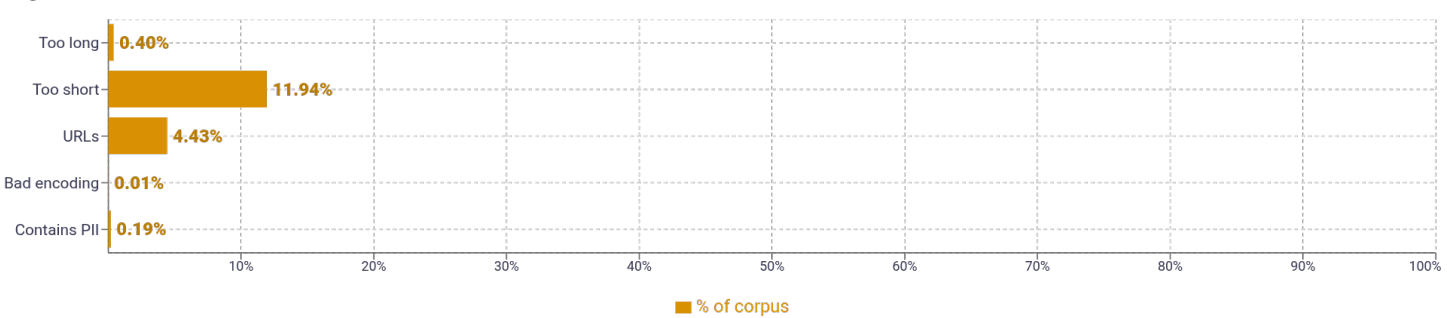
Frequent n-grams

SIZE	N-GRAMS	
1	aku | 39,433,532 tak | 35,808,314 nak | 31,998,633 ni | 31,796,139 dia | 31,574,977	⧉
2	a comment | 3,041,912 post a | 2,987,202 terima kasih | 2,922,212 laman web | 1,746,828 ibu bapa | 1,603,209	⧉
3	post a comment | 2,985,169 kah kah kah | 490,065 perintah kawalan pergerakan | 248,309 tepat pada masanya | 240,895 nabi muhammad saw | 218,028	⧉
4	assalamualaikum dan salam sejahtera | 123,213 ibu pejabat polis daerah | 93,727 tular di media sosial | 89,077 komen dan share ya | 83,922 bayi yang baru lahir | 78,056	⧉
5	lupa komen dan share ya | 76,587 jangan lupa komen dan share | 75,750 nik abdul aziz nik mat | 66,294 view this post on instagram | 56,082 komen yang diberikan oleh pembaca | 53,385	⧉

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				