

General overview

Corpus	Date	Language
hplt-v3-hin_Deva	9/18/2025	Hindi (hi)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
36,325,196	563,360,834	398,763,470 (70.78 %)	22B	99,222,596,922	230.32 GB

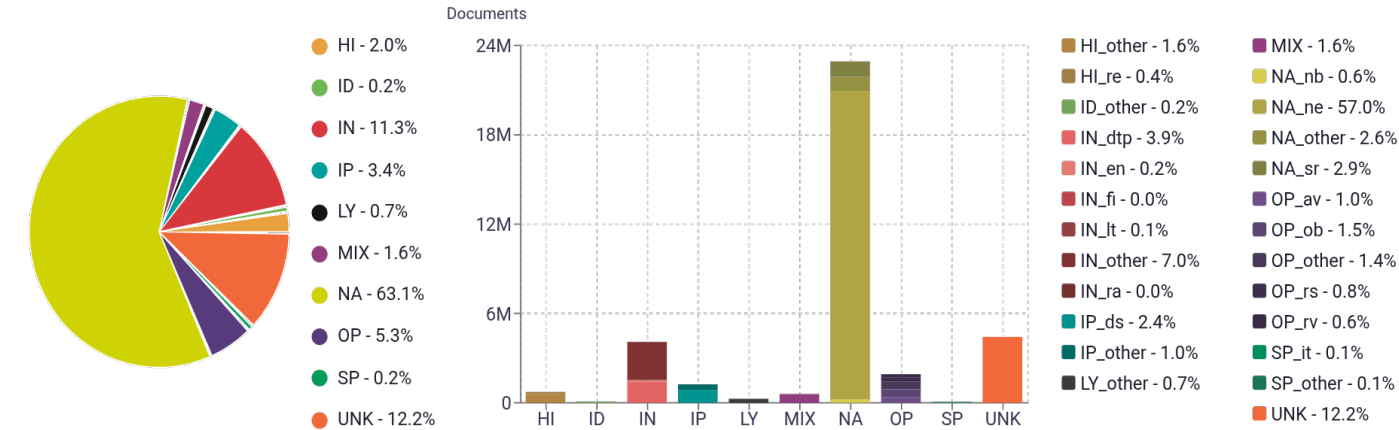
Top 10 domains

Domain	Docs	% of total
indiatimes.com	505K	1.39%
blogspot.com	493K	1.36%
jagran.com	469K	1.29%
bhaskar.com	418K	1.15%
patrika.com	357K	0.98%
news18.com	333K	0.92%
amarujala.com	288K	0.79%
punjabkesari.in	258K	0.71%
livehindustan.com	253K	0.70%
india.com	230K	0.63%

Top 10 TLDs

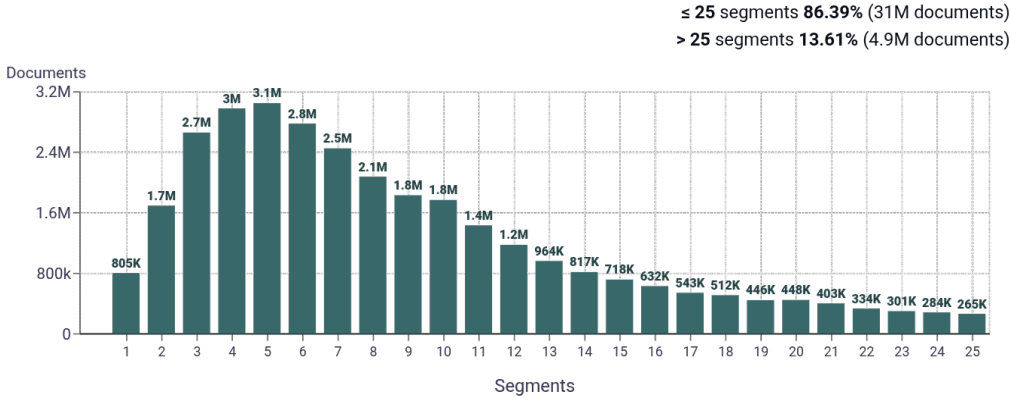
Domain	Docs	% of total
com	26M	71.52%
in	6.2M	16.93%
org	905K	2.49%
co.in	526K	1.45%
net	408K	1.12%
page	377K	1.04%
news	245K	0.67%
co	162K	0.44%
ru	123K	0.34%
xyz	121K	0.33%

Register labels

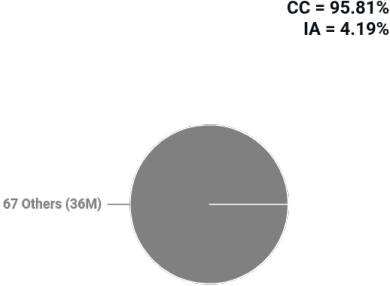


MT:7.0% | 2.5M Documents

Documents size (in segments) ⓘ

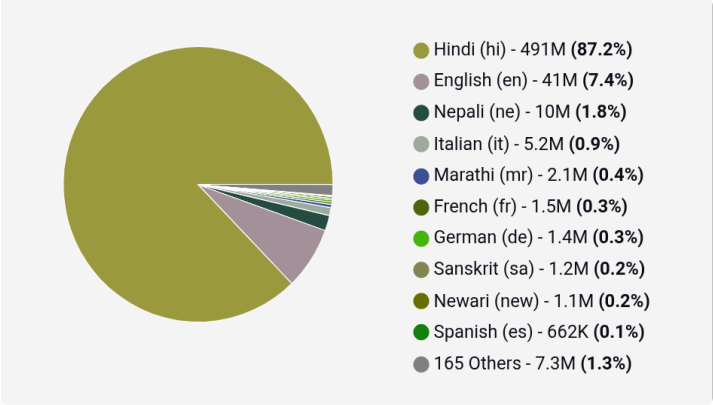


Document collections

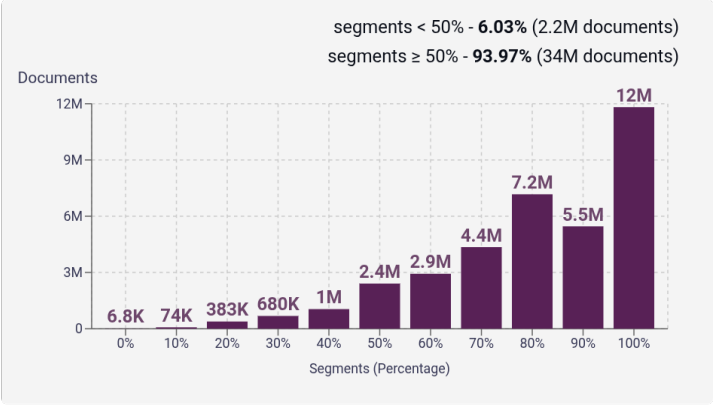


Language Distribution

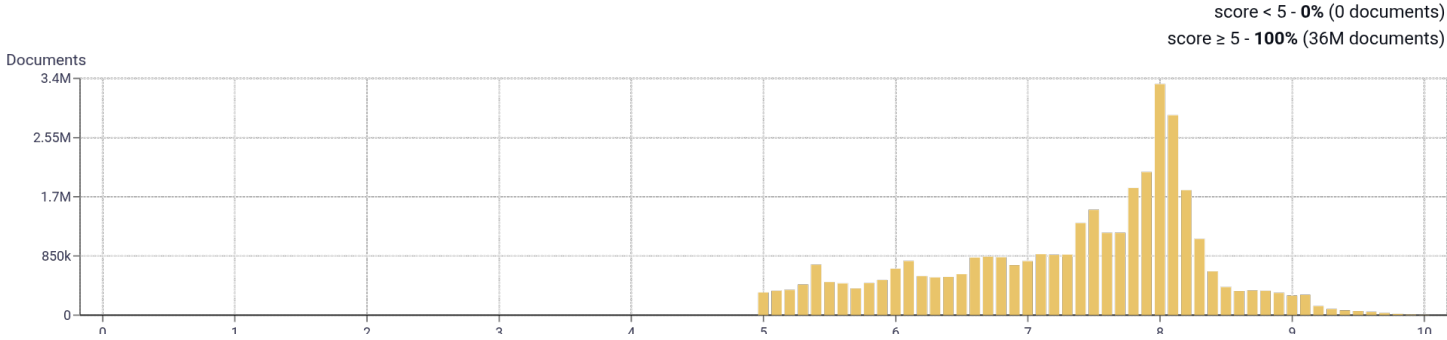
Number of segments in the Hindi (hi) corpus



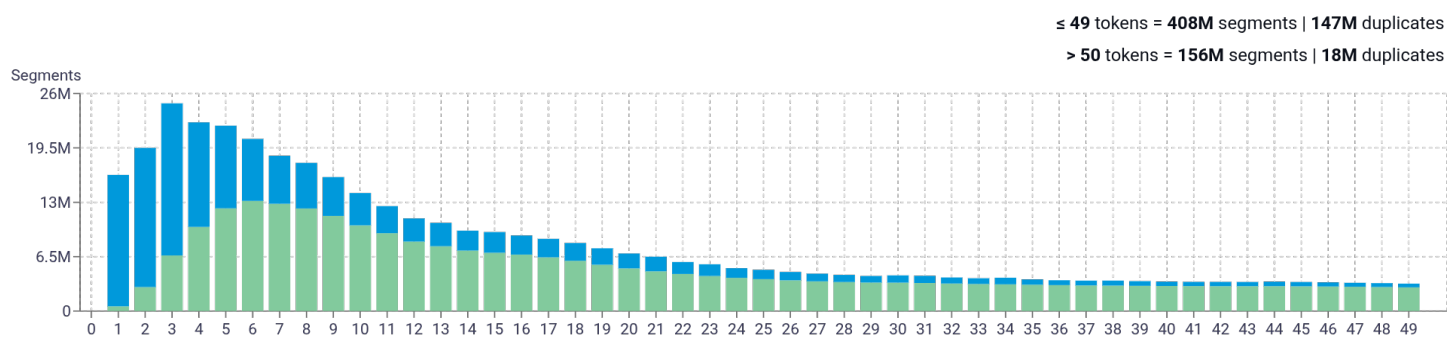
Percentage of segments in Hindi (hi) inside documents



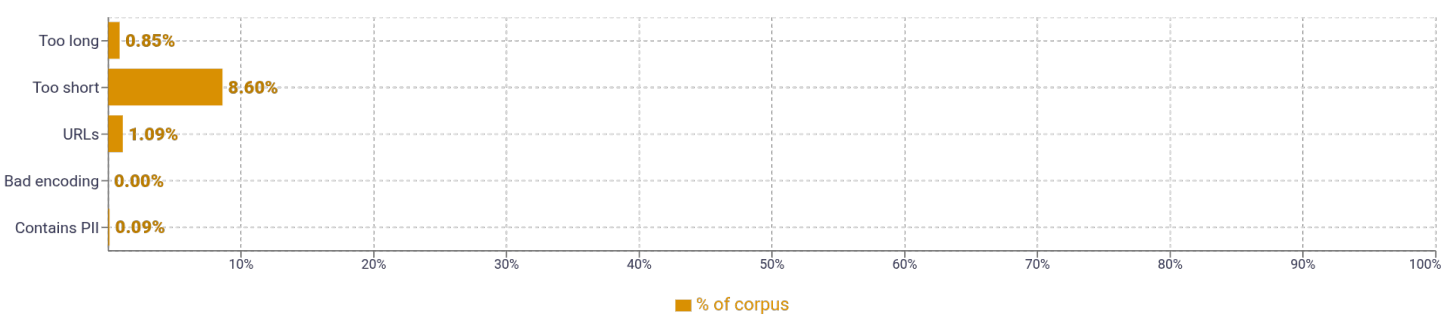
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	गई 34,305,856लेकिन 34,034,332रही 32,714,870पहले 32,086,004रूप 31,435,687	
2	दी गई 4,979,599बता दें 4,911,200किए गए 4,564,097सोशल मीडिया 4,551,281in hindi 4,195,761	
3	कम से कम 2,268,850मौत हो गई 1,711,697प्रधानमंत्री नरेंद्र मोदी 1,603,645योजना के तहत 1,437,433बारे में जानकारी 966,431	
4	अस्पताल में भर्ती कराया 364,708रूप में जाना जाता 346,078सोशल मीडिया पर वायरल 332,837कम करने में मदद 293,845नाम से जाना जाता 292,317	
5	लोगों की मौत हो गई 339,561नाम से भी जाना जाता 212,704जानकारी के लिए बता दें 167,090जान से मारने की धमकी 158,412view this post on instagram 155,315	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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				