No	Dataset name	Subset(s)	Dataset description	Dataset URL	HuggingFace URL	Dataset language(s)	Dataset collection region	Dataset task(s)	Dataset modality	Dataset domain(s)	Dataset license	Dataset annotation collection style	Dataset annotation validation style	Number of annotators per sample	Inter-annotator agreement
1	Abui WordNet		A small fully hand	https://github.con		abz	N/A	Wordnet	Language	Dictionary	Creative Commo	Machine-generat	Manual (full)		
2	AC-IQuAD	simple complex	This is an automa (question consist	https://www.kagg		ind	N/A	Question Answer	Language	Wikipedia	Creative Commo	Crawling	Automatic & Man		
3	AIFORTHAI - LotusCorpus		The Large vOcat	https://github.con		tha	Thailand	Automatic Speec	Language Speech	Multi-domain	Creative Commo	Expert-generatec	Manual (full)		
4	ALICE-THI	THI-C68 THI-D10	ALICE-THI is a T	https://www.ai.ru		tha	Thailand	Optical Characte	Language Vision	General	Unknown (unkno	Crowdsourced	Manual (full)		
5	AlloVera	javanese tagalog vietnamese	AlloVera, which p	https://github.con		jav tgl vie	N/A	Automatic Speec	Speech	Mixed	MIT (mit)	Crowdsourced	None		
6	Alorese Collection		Alorese Collectio	https://hdl.handle		aol	r Besar Village, A r Kecil Village, Al ulolong Village, Al naseli Village, Par ndai Village, Pan	Language Model Automatic Speec	Language Vision Speech	Conversational Mixed Culture & heritage Folklore Experiment or Task Wordlist Stimuli	Unknown (unkno	Expert-generatec	None		
7	ALT Burmese Treebank		A 20,000-senten	https://zenodo.or		mya	Myanmar	Constituency Par	Language	News articles	Creative Commo	Expert-generatec	None		
8	AM2iCo	Indonesian	We present AM2I	https://github.con		ind eng		Word Sense Disa	Language	Wikipedia	Creative Commo	Crawling Expert-generatec			
9	AMADI_LontarSet Binarization of P		AMADI_LontarSe	http://amadi.univ-		ban	Bali, Indonesia	Optical Characte	Language Vision	Culture & heritage	Unknown (unkno	Expert-generatec Machine-generat	Manual (full)		
10	AMADI_LontarSet Isolated Charact		AMADI_LontarSe	http://amadi.univ-		ban	Bali, Indonesia	Optical Characte	Language Vision	Culture & heritage	Unknown (unkno	Expert-generatec Machine-generat	Manual (full)		
11	AMADI_LontarSet Query-by-Examp		AMADI_LontarSe	http://amadi.univ-		ban	Bali, Indonesia	Optical Characte	Language Vision	Culture & heritage	Unknown (unkno	Expert-generatec Machine-generat	Manual (full)		
12	Amanatun wordlist	OZ_20220704-W OZ_20220704-WO	A wordlist of 122	https://catalog.pa		aoz	Oekiu	Word lists	Language Speech	atic Speech Reco Wordlist	Creative Commo	Expert-generatec	Automatic & Man		
13	ara-close-lange		The dataset cont The data from th	https://github.con		bcl ceb	Philippine	Readibility Asses	Language	Books	Creative Commo	Expert-generatec	None		
14	Asian Language Treebank Parallel		The ALT project a	https://www2.nict		khm lao mya ind fil zlm tha	N/A	Machine Translat	Language	News	Creative Commo	Expert-generatec	None		
15	Asian Signbank		The Asian Signba	http://cslds.org/a		hos inl	N/A	Machine Translat Sign Language R	Language Vision	Conversational	Other (other)	Expert-generatec	None		
16	ASR-IBSC: AN IBAN SPEECH CO		This package co	https://github.con		iba	Malaysia	Text-To-Speech S Automatic Speec	Speech	News	Creative Commo	Expert-generatec	None		
17	ASR-INDOCSC: AN INDONESIAN		This open-source	https://magichub.		ind	Asia	Text-To-Speech S Automatic Speec	Speech	Sports Food Movies/TV Family Holiday Friends	Creative Commo	Expert-generatec	None		
18	ASR-MALCSC: MALAY CONVERS		This open-source	https://magichub.		zlm	Johor, Malaysia	Text-To-Speech S Automatic Speec	Speech	Health Culture & heritage Sports Food Holiday Work Telecom	Creative Commo	Expert-generatec	None		
19	ASR-SINDODUSC: A SCRIPTED I		This open-source	https://magichub.		ind	Jakarta Bali Sulawesi East Java	Text-To-Speech S Automatic Speec	Speech	General	Creative Commo	Expert-generatec	None		
20	ASR-SMALDUSC: A SCRIPTED M		This open-source	https://magichub.		zlm	LANGOR, Malay	Automatic Speec Text-To-Speech S	Speech	General	Creative Commo	Expert-generatec	None		
21	ASR-STIDUSC: A SCRIPTED THA		This open-source	https://magichub.		tha	Bangkok, Thailand	Text-To-Speech S Automatic Speec	Speech	General	Creative Commo	Expert-generatec	None		
22	Audio Keyword Spotting	ind	This dataset is a	https://huggingfa	https://huggingfa	ind	N/A	Automatic Speec	Language Speech	Mixed	Creative Commo	Machine-generat	None		
23	Automated Similarity Judgment Pro		The database of	https://asjp.clld.o		ran khb tyr tlt sqq duq pgu tzn mlf dbe blw oni ktv xmz cbn sne ebk itx oia mqf gbi dms kgq mzt pll xxt etz tti mgf awr llm kiy zbw hmu cwg suc khp kzm iwo pne hil blf bdw btq lbt msl kkb wad ssq alo tyz kti bti bzp vie ssb bna kvr sml mxn lbw nev ksx pcb tfo mhe spr plc phq wnk mnx nec vto xsb bkr jah sbo sdo bvk kvw bhz bsb prh bps tom lpe lbk kja toy san snv ubl ahh sib kyp pku rth dmy kkh mhz kyi nun tlv kwr duo emb slz kwe jmn akb tih tlk brv awy ttd hbu khm kyn tiy ptu blj tmo lbl mso rac sou wod tmj lnd jbr aoz mhs sxm afz	N/A	Word lists	Language	General Mixed Culture & heritage Dictionary Academic Uncategorized wordnet Bilingual Dictionary Wordlist	Creative Commo	Expert-generatec Expert-translated	None		
24	Aya Collection - Translated	Adversarial QA CNN/DailyMail dolly Flan COPA Flan CoT Flan GEM Wiki Flan Lambada Flan Unified QA HotpotQA Joke Explanation Mintaka MLQA NQ-Open PAWS PIQA SODA Wiki-split WikiQA Xlel WD	The Aya Collectio	https://huggingfa		ceb tha mya zsm jav ind vie sun ace bjn khm lao min	Worldwide	Instruction Tuning	Language	General	Apache license 2	Machine-translat	None		
25	Aya Dataset		The Aya Dataset	https://huggingfa	https://huggingfa	tgl ceb tha mya zsm jav ind vie sun tam	Global	Instruction Tuning	Language	General Mixed Multi-domain	Apache license 2	Crowdsourced	None		
26	Aya Evaluation Suite	dolly_mt_ceb dolly_mt_sun dolly_mt_jav dolly_mt_zsm dolly_mt_ind dolly_mt_ace dolly_mt_vie dolly_mt_lao dolly_mt_min dolly_mt_mya dolly_mt_bjn dolly_mt_tam dolly_mt_khm dolly_mt_tha	Aya Evaluation S	https://huggingfa		ceb tha mya zsm jav ind vie sun ace bjn khm lao min	Worldwide	Instruction Tuning	Language	General	Apache license 2	Machine-translat	None		
27	Bactrian-X	Burmese Tagalog Indonesian Khmer Thai Vietnamese	The Bactrain-X d	https://github.con	https://huggingfa	mya tgl ind khm tha vie	Myanmar Philippines	Question Answer Summarization Relation Extractio Text Classificatio	Language	Mixed Multi-domain Wikipedia	Creative Commo	Machine-translat	Automatic & Man		0.7
28	BalitaNLP		BalitaNLP is a m	https://github.con		fil	Philippines	Image-to-Text Ge Text-to-Image Ge	Language Vision	News articles	Unknown (unkno	Crawling	Manual (partial)		
29	Barasa		Barasa: Indonesi	https://github.con		ind		Wordnet	Language	General	MIT (mit)	Unknown			
30	BasahaCorpus		BasahaCorpus c	https://github.con		bto krj hil msb	N/A	Readibility Asses	Language	Books	Creative Commo	Crawling	None		
31	Baybayin	aybayin character Latin characters bayin diacritic ima	The Baybayin da	https://www.kagg		tgl	Philippines	Optical Characte	Language Vision	General	Creative Commo	Crawling	Automatic		
32	Beaye Lexicon	lexicon	The Beaye Lexic	https://github.con		day	Landak Region	Word lists	Language	General	Apache license 2	Expert-generatec	Manual (full)	3	
33	Belebele	ceb_Latn ilo_Latn ind_Latn jav_Latn kac_Latn khm_Khmr lao_Laoo mya_Mymr shn_Mymr sun_Latn tgl_Latn tha_Thai vie_Latn war_Latn zsm_Latn	Belebele is a mul	https://github.con	https://huggingfa	ceb ilo ind jav kac khm lao mya shn sun tgl tha vie war zsm	N/A	Question Answer	Language	General	Creative Commo	Crowdsourced	Manual (partial)	2	
34	BEST	BEST 2010	The Benchmark f	https://github.con		tha	Thailand	Named Entity Re Statement Taggir	Language	Article Encyclopedia News Novel	Creative Commo	Expert-generatec	None		
35	BEST2019-Handwritten	68PersonsBmp ST200-1 ST200-2 ST200-3 WD200-1 WD200-2 WD200-3 WD200-4	This dataset cont	https://thailang.n		tha	Thailand	Optical Characte	Language Vision	Mixed	Unknown (unkno	Crawling Crowdsourcing	None		
36	Bhinneka Korpus		The Bhinneka Ko	https://github.con		day mak aoz mkn abs ind	Central Timor Re Kupang Regency Landak Regency Ambon Regency Makasar Regency	Machine Translat	Language	Conversational General	Apache license 2	Crowdsourced Expert-generatec	Manual (full)		
37	Bible Corpus	Vietnamnese Bible Tagalog Bible Indonesian Bible Galela Bible Thai Bible Cebuano Bible Uma Bible	A multilingual pa	https://github.con		vie tgl ind tha gbi ceb ppk	N/A	Machine Translat	Language	Religion	Creative Commo	Crawling Expert-generatec	None		
38	BKD-Prosub		BKD-Prosub is a	https://babyai-hu		tha	Thailand Japan	Coreference Res	Language	Books Movies/TV	Apache license 2	Expert-generatec	Manual (full)		

No	Train split data size	Validation split data size	Test split data size	No split data size	Data size unit	Total data size	Dataset provider	Dataset paper title	Dataset paper URL	Dataset or dataset paper publish year	Publication venue	Dataset access	Derived from...	Does the data have PII (personally identifiable information)?	Does the data have sensitive content/information?	Submission Date	Approval Status	Dataloader name	GitHub Issue URL
1				3606	instances	3606	Palacký Universi	Abui Wordnet: Us	https://aclantholo	2022	ACL Workshop	Free		No	No	2023-11-24 18:2	Approved	abui_wordnet	https://github.con
2	234 171	0 0	219 72		instances	696	Universitas Indor	AC-IQuAD: Autor	https://link.spring	2023	Language Resou	Free		No	No	2024-02-22 08:3	Approved	ac_iquad	https://github.con
3	3007	500	500		sentences	4007	TMEC Prince of Songkla Mahanakorn Univ	LOTUS: Large vC	https://doi.org/10.	2005	NSTDA Annual C	Free		No	No	2024-02-15 12:5	Approved	tha_lotus	https://github.con
4	13130 8055	0 0	1360 1500		images	24045	University of Gro	Recognition of ha	https://www.scien	2015	Engineering Appl	Free		No	No	2023-12-24 02:3	Approved	alice_thi	https://github.con
5				N/A (see Table 1)	instances	0	Carnegie Mellon	AlloVera: A Multil	https://aclantholo	2020	LREC	Free		No	No	2024-04-01 11:2	Approved with notes	allovera	https://github.con
6				Contact me for d	hours	0	Leiden University	The Plural Word	https://scholarlyp	2016	Oceanic Linguisti	Free		Yes	Yes	2024-02-14 23:3	Approved	alorese	https://github.con
7	9187	0	6123		sentences	15310	NICT Japan	A Burmese (Myar	https://dl.acm.org	2020	TALLIP	Free		No	No	2023-11-01 12:5	Approved	alt_burmese_tree	https://github.con
8	1,589	500	1,000		examples	3089	Public	AM2 ICO: Evalua	https://arxiv.org/p	2021		Free		No	No		NusaCrowd carry-over	id_am2ico	
9	150	N/A	50		images	10430	University of La F University Padjac Institute of Techn Universitas Pend	AMADI_LontarSe	https://hal.scienc	2016	ICFHR	Free		No	No	2023-12-20 02:3	Approved	amadi_lontarset_	https://github.con
10	±10000	N/A	±7000		images	0	University of La F University Padjac Institute of Techn Universitas Pend	AMADI_LontarSe	https://hal.scienc	2016	ICFHR	Free		No	No	2024-05-31 18:2	Approved	amadi_lontarset_	https://github.con
11	10100	N/A	130		images	10230	University of La F University Padjac Institute of Techn Universitas Pend	AMADI_LontarSe	https://hal.scienc	2016	ICFHR	Free		No	No	2024-05-31 18:1	Approved	amadi_lontarset_	https://github.con
12				1228	tokens	1228	OCSEAN	Amanatun wordli	https://www.ocse	2022	Paradisec	Free		No	No	2024-03-22 13:1	Approved with notes	amanatun_wordli	https://github.con
13	0	0	7200		documents	7200	National Universi	Automatic Reada	https://aclantholo	2023	ACL	Free	Let's Read Asia (	No	No	2023-11-01 08:4	Approved	ara_close	https://github.con
14	18088	1000	1018		sentences	20106	multiple institutio	Introduction of th	https://www2.nict	2016	O-COCOSDA	Free	Wikinews	No	No	2023-11-20 19:5	Approved	parallel_asian_tr	https://github.con
15				N/A	instances	0	The Chinese Uni	Sign Language D	https://www.sign-	2012	LREC	Free upon reque		No	No	2023-12-16 07:5	Approved	asian_signbank	https://github.con
16	6.8	N/A	1.2		hours	8	Grenoble Informa University Greno Faculty of Compu	Semi-supervised		2014	SLTU	Free		No	No	2024-02-11 07:1	Approved	asr_ibsc	https://github.con
17				4.54	hours	4.54	Beijing Magic Da	N/A		2021	No publication	Free		No	No	2024-02-11 06:2	Approved	asr_indocsc	https://github.con
18				5	hours	5	Beijing Magic Da	N/A		2021	No publication	Free		No	No	2024-02-11 08:3	Approved	asr_malcsc	https://github.con
19				3.5	hours	3.5	Beijing Magic Da	N/A		2021	No publication	Free		No	No	2024-02-11 06:3	Approved	asr_sindodusc	https://github.con
20				4.8	hours	4.8	Beijing Magic Da	N/A		2021	No publication	Free		No	No	2024-02-11 08:2	Approved	asr_smaldusc	https://github.con
21				4.56	hours	4.56	Beijing Magic Da	N/A		2021	No publication	Free		No	No	2024-02-11 08:3	Approved	asr_stidusc	https://github.con
22	575	83	81		utterances	739	SIL International	-	https://huggingfa	2022	No publication	Free	CommonVoice	Yes	No	2023-12-03 21:4	Approved	audio_keyword_s	https://github.con
23				36000	instances	36000	Max Planck Instit	N/A		2022	No publication	Free		No	No	2024-03-31 05:0	Approved	asjp	https://github.con
24				0	instances	0	Cohere for AI	Aya Dataset: An	https://arxiv.org/a	2024	Preprint	Free	dolly v2	No	No	2024-03-03 15:3	Approved	aya_collection_tr	https://github.con
25				N/A	prompt/completic	0	Cohere for AI	Aya Dataset: An	https://arxiv.org/a	2024	Preprint	Free		No	No	2024-02-14 06:0	Approved	aya_dataset	https://github.con
26			200 200 200 200 200 400 200 200 400 200 400 200 200 200		instances	3400	Cohere for AI	Aya Dataset: An	https://arxiv.org/a	2024	Preprint	Free	dolly v2	No	No	2024-03-03 15:2	Approved	aya_evaluation_	https://github.con
27	67000	0	0		sentence pairs	67000	The University O MBZUAI	Bactrian-X: Multil	https://arxiv.org/p	2023	Preprint	Free	alpaca-52k and c	No	No	2024-01-30 06:3	Approved	bactrian_x	https://github.con
28	262480	32806	32821		examples	328107	Angeles Universi	Transformer-base	https://storage.gc	2023	Proceedings of th	Free		No	Yes	2023-11-01 05:3	Approved	balita_nlp	https://github.con
29	16	0	0		MB	16.2	Ahmad Izzan, Ch			2015		Free					NusaCrowd carry-over	barasa	
30	7784	1113	2225		documents		National Universi University of Batt MBZUAI	BasahaCorpus: A	https://arxiv.org/a	2023	EMNLP	Free		No	No	2023-11-05 00:1	Approved	basaha_corpus	https://github.con
31				17000 18200 2000	images	37200	University of the	Optical character	https://peerj.com/	2021	PeerJ Computer	Free	Multiple datasets	No	No	2023-12-25 12:1	Approved	baybayin	https://github.con
32	6250	782	782		tokens	7814	Independent	Penyusunan Kor	https://drive.goog	2023	Kongres Bahasa	Free	NusaX Lexicon D	No	No	2023-10-31 10:2	Approved	beaye_lexicon	https://github.con
33	3000	500	500		instances	4000	Meta	The BELEBELE I	https://arxiv.org/p	2023	Preprint	Free	FLORES-200	No	No	2023-10-29 18:5	Approved	belebele	https://github.con
34	198 108 96 107	N/A	N/A		documents	509	Chulalongkorn U Kasetsart Univer NECTEC TSL	Thoughts on wor	http://pioneer.chu	2010	Proceedings of th	Free		No	No	2024-03-04 00:0	Approved with notes	best	https://github.con
35				68 25 22 32 38 54 32 32	documents	303	NECTEC	Thai Handwritten	https://link.spring	2019	International Wor	Free		Yes	No	2024-03-30 17:1	Approved	best2019	https://github.con
36	700	100	184		sentence pairs	984	Independent	Penyusunan Kor	https://drive.goog	2023	Kongres Bahasa	Free	Tatoeba Project	No	No	2023-10-30 16:3	Approved	bhinneka_korpus	https://github.con
37				31102 31102 29813 7750 31102 31113 7793	paragraphs	169775	University of Edir	A massively para	https://link.spring	2015	Language Resou	Free		No	Yes	2023-12-28 03:0	Approved	bible_corpus_chr	https://github.con
38				10360 12624 668	tokens	23652	Tokyo University Musashino Unive	Corpus Developr	https://www.virac	2023	International Con	Free		No	No	2024-03-29 11:1	Approved with notes	bkd_prosub	https://github.con

No	Dataset name	Subset(s)	Dataset description	Dataset URL	HuggingFace URL	Dataset language(s)	Dataset collection region	Dataset task(s)	Dataset modality	Dataset domain(s)	Dataset license	Dataset annotation collection style	Dataset annotation validation style	Number of annotators per sample	Inter-annotator agreement	Train split data size	Validation split data size	Test split data size	No split data size	Data size unit	Total data size	Dataset provider	Dataset paper title	Dataset paper URL	Dataset or dataset paper publish year	Publication venue	Dataset access	Derived from...	Does the data have PII (personally identifiable information)?	Does the data have sensitive content/information?	Submission Date	Approval Status	Dataloader name	GitHub Issue URL
39	Bloom-captioning	abc ahk bfn bjn bkx brb brv bya bzi ceb cgc cmo ddg dmg dnw dtp dtr enc fil gal hil hre hro idt ilo ind jra kak khb khm kqr krr ksw lhu llg lsi lwl mdr mgm mhx mkz mnw mqj mry msb mya nod nst nxa nxl pag pce pdu pea pmf sea sgd shn sml snl tdt tet tha tkd tnt tom tpu vie war wms wnk xmm yet zlm	This is a Bloom L	https://huggingfa	https://huggingfa	abc ahk bfn bjn bkx brb brv bya bzi ceb cgc cmo ddg dmg dnw dtp dtr enc fil gal hil hre hro idt ilo ind jra kak khb khm kqr krr ksw lhu llg lsi lwl mdr mgm mhx mkz mnw mqj mry msb mya nod nst nxa nxl pag pce pdu pea pmf sea sgd shn sml snl tdt tet tha tkd tnt tom tpu vie war wms wnk xmm yet zlm	N/A	Image-to-Text Ge	Language Vision	Books	Creative Commo	Expert-generatec	None						8 907 97 62 74 330 7 7 497 2953 1638 199 118 10 11 161 10 5 560 96 26 35 40 64 27 2177 1423 1416 167 311 28 353 19 349 18 202 46 8 107 945 152 10 17 68 6 421 29 23 101 34 7 314 10 20 12 33 42 5 25 6 191 65 3023 45 17 17 521 22 154 17 18 18 1 102	images	21057	SIL International	Bloom Library: M	https://aclantholo	2022	EMNLP	Free upon reque	Bloom Library	No	No	2023-12-02 00:4	Approved	bloom_captioning	https://github.com
40	Bloom-LM	ahk bfn bkx brb bzi ceb cgc cmo ddg dtp fil gal idt ind jra kak khb khm krr lhu lsi mgm mhx mkz mya nxa pce tdt tet tha tpu war zlm	This version of th	https://huggingfa	https://huggingfa	ahk bfn bkx brb bzi ceb cgc cmo ddg dtp fil gal idt ind jra kak khb khm krr lhu lsi mgm mhx mkz mya nxa pce tdt tet tha tpu war zlm	N/A	Language Model	Language	Books	Other (other)	Crawling	None			81 11 8 18 53 335 158 16 14 13 55 11 8 208 112 156 25 28 24 32 21 12 79 16 31 12 30 22 8 228 24 16 8	10 1 1 2 7 42 20 2 2 2 7 1 1 26 14 20 3 4 3 4 3 2 10 2 4 1 4 3 1 29 3 2 1	10 1 1 2 6 41 19 2 1 1 6 1 1 25 13 19 3 3 3 4 2 1 9 2 3 1 3 2 1 28 2 2 1		instances	2329	SIL International	Bloom Library: M	https://aclantholo	2022	EMNLP	Free upon reque		No	No	2023-12-02 06:0	Approved	bloom_lm	https://github.com
41	Bloom-speech	bjn bzi ceb ind jra kqr mya tgl	This version of th	https://huggingfa	https://huggingfa	bjn bzi ceb ind jra kqr mya tgl	N/A	Automatic Speec Text-To-Speech S	Language Speech	Books	Other (other)	Crawling	Manual (full)			80 1363 553 0 203 0 321 352	20 50 50 0 50 0 50 48	50 157 67 14 50 18 50 50		instances	3596	SIL International	Bloom Library: M	https://aclantholo	2022	EMNLP	Free upon reque		No	No	2023-12-02 05:4	Approved	bloom_speech	https://github.com
42	Bloom-vist	abc ahk bfn bjn bkx brb brv bya bzi ceb cgc cmo ddg dmg dnw dtp enc fil hil hro idt ilo ind jra kak khb khm kqr krr ksw lhu lsi lwl mdr mgm mhx mkz mry msb mya nod nxa nxl pag pce pea pmf psp sea sgd sml snl tdt tet tha tkd tpu war wms yet yin zlm	BLOOM VIST is a	https://huggingfa	https://huggingfa	abc ahk bfn bjn bkx brb brv bya bzi ceb cgc cmo ddg dmg dnw dtp enc fil hil hro idt ilo ind jra kak khb khm kqr krr ksw lhu lsi lwl mdr mgm mhx mkz mnw mqj mry msb mya nod nxa nxl pag pce pea pmf psp sea sgd sml snl tdt tet tha tkd tpu war wms yet yin zlm	Unknown	Visual Storytelling	Language Vision	Books	Creative Commo	Expert-generatec	None						1 55 1 5 3 22 1 1 66 394 182 20 2 1 1 15 1 34 3 1 1 2 232 56 195 31 26 1 29 1 9 22 6 1 1 51 4 6 1 11 1 1 2 1 9 1 1 6 1 8 2 1 10 3 275 2 28 17 1 1 1 10	documents	1877	SIL International	Bloom Library: M	https://aclantholo	2022	EMNLP	Free upon reque		No	No	2023-11-26 16:0	Approved	bloom_vist	https://github.com
43	BM-PT3	BM-A-PT3 BM-B-PT3	This test is for 15	https://github.com		zlm	Malaysia	Commonsense R	Language	School Exam	Unknown (unkno	Crowdsourced	Manual (full)						54 11	instances	65	Independent	NA		2023	NA	Free		No	No	2024-03-14 16:5	Approved	bm_pt3	https://github.com
44	BRCC		The Bahasa Roja	https://data.depo		zlm eng cmn	bahasarojak/resc	Language Model	Language	Wikipedia	Unknown (unkno	Crawling Machine-generat	Manual (partial)						2640702	sentences	2640702	National Central Academia Sinica	BRCCandSentiB	https://aclantholo	2022	COLING	Free		No	No	2024-02-27 10:4	Approved	brcc	https://github.com
45	Bud500		Bud500 is a dive	https://huggingfa	https://huggingfa	vie	Vietnam	Automatic Speec	Language Speech	Mixed	Apache license 2	Crawling Crowdsourced	Manual (full)			634158	7500	7500		utterances	649158	VietAI Research	Bud500: A Comp	https://github.com	2024	No publication	Free		No	No	2024-03-16 11:4	Approved	bud500	https://github.com
46	Burapha-TH	character digit syllable	The dataset has	https://services.ir		tha	Thailand	Optical Characte	Language Vision	General	Unknown (unkno	Expert-generatec	Manual (full)			N/A	N/A	N/A		images	0	Burapha Univers	Burapha-TH: A M	https://www.mdpi	2022	MDPI	Free		No	No	2023-11-19 14:4	Approved	burapha_th	https://github.com
47	CASA		CASA is an aspe each label repres	https://github.com	https://huggingfa	ind		Aspect Based Se	Language	reviews	Creative Commo	Crawling Expert-generatec				810	90	180		sentences	1080	Christian Wibiso	Aspect detection classification usir	https://ieeexplore	2018		Free	Online Platform				NusaCrowd carry-over	casa	
48	CC-100	id jv km lo ms my my_zaw su th tl vi	This corpus comp	https://data.statm		mya lao khm tgl vie tha zlm ind jav sun	N/A	Language Model	Language	Web	MIT (mit)	Crawling	None						36000 37 153 63 2100 46 178 15 8700 701 28000	MB	75993	Meta	Unsupervised Cr	https://aclantholo	2020	ACL	Free	CCNet	No	No	2023-11-28 19:0	Approved	cc100	https://github.com
49	CC-Aligned	id_ID su_ID th_TH vi_VN ms_MY lo_LA km_KH my_MM my_MM_zaw cx_PH wy_PH	CCAligned consi	https://www2.stat	https://huggingfa	ind sun tha vie zlm lao khm mya ceb war eng	N/A	Machine Translat	Language	Web	Unknown (unkno	Crawling	Manual (partial)						1600 37 2000 1500 875 71 93 69 8.8 20 0.753	MB	6,274.55	Meta	CCAligned: A Ma	https://aclantholo	2020	EMNLP	Free	Common Crawl	No	No	2023-11-29 17:0	Approved	cc_aligned_doc	https://github.com
50	CC-Aligned Sentence Pairs	id_ID jv_ID su_ID th_TH vi_VN ms_MY lo_LA km_KH my_MM cx_PH	This dataset cont	https://www2.stat		ind jav sun tha vie zlm lao khm mya ceb	N/A	Machine Translat	Language	Web	Unknown (unkno	Machine-generat	None						759 24 19 533 459 216 6.9 21 15 9.1	MB	2,062	Meta	Low-Resource C	https://aclantholo	2020	WMT	Free	CC-Aligned	No	No	2023-11-29 17:1	Approved	cc_aligned_sent	https://github.com
51	CC100	Indonesian Javanese Sundanese	CC100 comprise	https://data.statm	https://huggingfa	ind sun jav		Language Model	Language	multi domain	Other (other)	Crawling				36,052	0	0		MB	36052	Kata.ai	Unsupervised Cr	https://aclantholo	2020		Free	Common Crawl	No	No		NusaCrowd carry-over	cc100	
52	CC3M-35L	fil ind tha vie	CC3M-35L is cre	CC3M-35L: https CC3M: https://ai.		fil ind tha vie	N/A	Image-to-Text Ge	Language Vision	Multi-domain	Creative Commo	Machine-generat	None			3318333 3318333 3318333 3318333	15840 15840 15840 15840	0		examples	13336692	Google	Crossmodal-360	https://aclantholo	2022	EMNLP	Free	Conceptual Capt	No	No	2023-11-15 21:0	Approved	cc3m_35l	https://github.com
53	CCMatrix		The CCMatrix da	https://opus.nlpl.	https://huggingfa	jav eng vie ind tgl mya zlm	US	Language Model Machine Translat	Language	Web	BSD license fam	Crawling	Automatic						N/A	sentence pairs	0	Meta	CCMatrix: Mining	https://aclantholo	2021	ACL	Free		No	No	2023-12-31 16:3	Approved with notes	ccmatrix	https://github.com
54	CebuaNER		The CebuaNER	https://github.com		ceb	N/A	Named Entity Re	Language	News	Creative Commo	Crawling Expert-generatec	Manual (full)	2	0.9315	622	77	77		sentences	4258	National Universi University of Batt Siliman Universit Bicol University	CebuaNER: A Ne	https://arxiv.org/a	2023	PACLIC	Free		No	No	2023-11-04 23:5	Approved	cebuaner	https://github.com
55	Cendana		Cendana is a ling	https://github.com	-	ind		Constituency Par	Language	chat	GNU General Pu	Expert-generatec				552	0	0		sentences	552	Ferdiant Joshua,	Building Cendana	https://core.ac.uk	2019		Free	Traveloka Conve	No	No		NusaCrowd carry-over	cendana	
56	chatgpt-malaysian-open-qa	Common Crawl Q/ hansard-qa.jsonl wikipedia-qa.jsonl	Synthetic Malays Generated using - common-crav - hansard-qa.js - wikipedia-qa.	https://huggingfa	https://huggingfa	zlm	Malaysia	Question Answer	Language	Web	Creative Commo	Machine-generat	Automatic						69829 42368 44923	examples	157120	Independent Mesolitica	NA		2024	No publication	Free	MS Wikipedia, M	No	No	2024-03-15 09:5	Approved	chatgpt_malaysia	https://github.com
57	CLIRMatrix		CLIRMatrix is a r	https://github.com		tgl ilo min jav sun ceb vie tha	N/A	Text Retrieval	Language	Wikipedia	Unknown (unkno	Machine-generat	Automatic			Contact me for d	Contact me for d	Contact me for d		documents	0	Johns Hopkins U	CLIRMatrix: A ma	https://aclantholo	2020	EMNLP	Free	Wikipedia	Yes	No	2024-02-02 14:3	Approved with notes	clir_matrix	https://github.com

No	Dataset name	Subset(s)	Dataset description	Dataset URL	HuggingFace URL	Dataset language(s)	Dataset collection region	Dataset task(s)	Dataset modality	Dataset domain(s)	Dataset license	Dataset annotation collection style	Dataset annotation validation style	Number of annotators per sample	Inter-annotator agreement	Train split data size	Validation split data size	Test split data size	No split data size	Data size unit	Total data size	Dataset provider	Dataset paper title	Dataset paper URL	Dataset or dataset paper publish year	Publication venue	Dataset access	Derived from...	Does the data have PII (personally identifiable information)?	Does the data have sensitive content/information?	Submission Date	Approval Status	Dataloader name	GitHub Issue URL
58	CMU Wilderness Multilingual Spee		The CMU Wilden	http://festvox.org		mhx ifk tlb nod ilo frd cgc tha cfm bgr blt atq dtp cmr amk ptu jav lsi nij mhy acn prf alj lnd kzf pww sda mbb ify mbt iba pse kje gbi mog alp twb law dni ahk rej bcl nlc plw zyp lew mad txa bpr min kne agn mqj itv gor bts twu mwv sml npy khm sas krj ury obo kqe mrw ifb mvp cmo por xsb ljp bru ban ind cnk sgb mak nia sun hnn ceb btd lao pam kac ifa blz bps ctd mnb pmf hil sxn bep ppk mej ace ifu tgl lex vie btx lhu pag xmm bhz tby	N/A	Automatic Speec Text-To-Speech S	Language Speech	Religion	Unknown (unkno	Crawling Expert-generatec Machine-generat	Automatic						20	utterances	1389443	Carnegie Mellon	CMU Wilderness	https://ieeexplore	2019	ICASSP	Free	Bible.is	No	Yes	2023-12-30 15:3	Approved	cmu_wilderness	https://github.com
59	COCO Captions ID		COCO Captions	https://drive.goog		ind		Image-to-Text Ge	Language Vision	General	Creative Commo	Machine-translat				113,287	5,000	5,000		sentences	123287	Microsoft, Ray Ar	Pembangkitan De	https://digilib.itb.a	2019		Free		No	No		NusaCrowd carry-over	coco_caption_id	
60	COCO-35L	fil ind tha vie	COCO-35L is a n	https://google.gitl		fil ind tha vie	N/A	Image-to-Text Ge	Language Vision	Multi-domain	Creative Commo	Machine-generat	None			113287 113287 113287 113287	5000 5000 5000 5000	0		examples	473148	Google	Crossmodal-3600	https://aclantholo	2022	EMNLP	Free	COCO Captions	No	No	2023-11-15 21:07	Approved	coco_35l	https://github.com
61	Code-mixed Sentiment JV-ID	ent Analysis - Jav ent Analysis - Indo lachine Translatio	Dataset terdiri da	https://docs.goog		jav ind		Sentiment Analys Machine Translat	Language	Social media	Creative Commo	Expert-generatec				977	0	0		tweets	977	Universitas Indor	Code-mixed sent	https://iopscience	2021		Free	-	No	No		NusaCrowd carry-over	code_mixed_jv_i	
62	CodeSwitch-Reddit	eng_monolingual cs	This corpus cons	https://github.com		eng ind tgl	N/A	Language Model Code-switching I	Language	Social media	Unknown (unkno	Crowdsourced	Manual (full)	3					211707 135703	examples	347410	University of Toro	CodeSwitch-Red	https://aclantholo	2019	EMNLP	Free		Yes	No	2024-01-11 11:30	Approved	codeswitch_redd	https://github.com
63	CommonVoice12.0	tha vie ind cnh	The Common Mc	https://commonv	https://huggingfa	tha vie ind cnh	N/A	Automatic Speec	Language Speech	General	Creative Commo	Crowdsourced	Manual (full)			32148 2475 5040 817	10964 285 3288 761	10964 1302 3647 763		utterances	72454	Mozilla	Common Voice: A	https://aclantholo	2020	LREC	Free upon reque		No	No	2023-11-22 01:26	Approved	commonvoice_12	https://github.com
64	CoNLL 2017 Shared Task Universa	UDP-Indonesian UDP-Vietnamnese	CoNLL Universal	https://lindat.mff.c		ind	Indonesia	Dependency Par	Language	Mixed	Creative Commo	Crawling Machine-generat	Automatic						5205000 4066000	tokens	9271000	Charles Universit	CoNLL 2017 Sha	https://aclantholo	2017	CoNLL	Free		No	No	2024-04-01 02:15	Approved	conll2017_udp	https://github.com
65	COPAL	test test_colloquial	a novel Indonesia	https://huggingfa	https://huggingfa	ind	Jakarta	Commonsense R	Language	Culture & heritage	Creative Commo	Expert-generatec	Automatic & Man	1	3				559	sentences	559	Independent MBZUAI Pitik	COPAL-ID: Indon	https://arxiv.org/p	2023	Preprint	Free		No	No	2023-11-08 07:34	Approved	copal	https://github.com
66	CORD	Indonesian	In this paper, we Furthermore, it a	https://drive.goog	https://huggingfa	ind		Optical Characte	Language	Receipt	Creative Commo	Crawling Expert-generatec				800	100	100		receipts	1000	Universal Depen	CORD: A Consoli	https://openreview	2019		Free		No	No		NusaCrowd carry-over	cord_v2	
67	Corpus Crawler		Primarily a crawle	https://github.com		blz cnh mvp tlb dtp mqj mog kne dnw mta ppk cbl lex bps nlc zyp bpr msm mej hnn shn lew sbl sgb czt tby ifa kqe mnw itv ahk due ury ptu atb bgr dtb amk yva lnd mtj bts mwv iba kyu pww cgc hil loe tih nij agn bhw bgz rej wrs ceb sas lis cnk mbt tyz duo mak gor prf kje kxm kgr ccp cfm yli aoz sxn iry lhu txa obo pmf ifb bzi atq ifk xsb pez btx gbi ljp mnb bkl akb alp npy ifu mqn pag bku btd mbb mrw alj bru bvz zlm min nia twu otd aaz law	N/A	Word lists	Language	Religion	Apache license 2	Crawling	None						N/A	words	0	Google	N/A		2017	No publication	Free		No	No	2023-12-23 20:5	Approved with notes	corpus_crawler	https://github.com
68	Corpus of Singapore English Mess		The CoSEM data The dataset is in	https://github.com		eng	Singapore	Language Model	Language	Social media	Creative Commo	Crowdsourced	None						1000	utterances	1000	The Chinese Uni University of Mich	The Corpus of Si	https://onlinelibra	2021	World Englishes	Free		No	No	2023-11-14 15:14	Approved	cosem	https://github.com
69	CoVoST 2	Indonesian English	With the aim to f	https://github.com	https://huggingfa	ind eng		Speech-to-Text T	Speech Language	Banking General	Creative Commo	Crawling Expert-translated				1	1	1		hours	3	Computer Scienc	CoVoST 2 and M	https://arxiv.org/p	2020		Free	CoVo, Tatoeba	No	No		NusaCrowd carry-over	covost2	
70	CreoleRC		CreoleRC is a su	https://github.com		cbk	N/A	Relation Extractio	Language	General Wikipedia	Creative Commo	Crawling	Manual (full)						97	relations	97	Aalborg Universit	CreoleVal: Multili	https://arxiv.org/a	2023	Preprint	Free	Wikipedia	No	No	2023-12-23 18:22	Approved	creole_rc	https://github.com
71	Cross-Lingual Data Augmentation f	th_fasttext_aug th_llm_gec_aug llm_paraphrase_a th_ltw2v_aug th_qcpg_0.2_aug cpg_0.2_llm_gec_ th_qcpg_0.5_aug cpg_0.5_llm_gec_ th_qcpg_0.8_aug cpg_0.8_llm_gec_ th_thai2fit_aug h_thai2trans_aug th_wordnet_aug en_aug en_llm_gec_aug llm_paraphrase_a en_qcpg_0.2_aug cpg_0.2_llm_gec_ en_qcpg_0.5_aug cpg_0.5_llm_gec_ en_qcpg_0.8_aug cpg_0.8_llm_gec_	This paper prese	https://github.com	https://huggingfa	tha eng	Thailand	Question Answer	Language	Banking Books Education General Multi-domain News articles Wikipedia Wikidata World Technology	MIT (mit)	Expert-generatec	Manual (full)						16980 16980	examples	373560	Independent VISTEC	Cross-Lingual Da	https://aclantholo	2023	EMNLP	Free	iapp-wiki-qa, tydi	No	No	2024-01-06 11:43	Approved	cross_lingual_au	https://github.com
72	Cross-lingual Outline- based Dialog	Indonesian	Cross-lingual Ou standing, dialogu diverse language	https://github.com		ind		Dialogue System	Language	multi-domain	Unknown (unkno	Machine-translat Expert-translated				0	92	102		dialogues	194	-	Cross-Lingual Di	https://arxiv.org/a	2022		Free	English Schema-	No	No		NusaCrowd carry-over	cod	
73	CrossSum	id-my id-vi my-id my-vi vi-id vi-my	This is a large-sc	https://drive.goog	https://huggingfa	ind vie mya	N/A	Cross-lingual Abs	Language	News articles	Creative Commo	Crowdsourced	Automatic & Man			615 3832 615 360 3832 360	76 479 76 45 479 45	76 479 76 45 479 45		article-summary p	12014	Bangladesh Univ Monash Universi University of Cali Swinburne Unive	CrossSum: Beyo	https://aclantholo	2023	ACL	Free	XL-Sum	No	No	2024-01-23 17:32	Approved	crosssum	https://github.com
74	CUB Bahasa		Semi-translated c	https://github.com		ind	N/A	Image-to-Text Ge	Language Vision	Birds	Unknown (unkno	Machine-generat	Manual (partial)			8855	2933	N/A		images	11788	Bandung Institute	Indonesian Text-t	https://arxiv.org/a	2023	Preprint	Free	CUB-200-2011 ((	No	No	2023-11-15 23:07	Approved	cub_bahasa	https://github.com
75	CulturaX	id jv km lo tl min my su th vi ms ceb war cbk bcl	CulturaX is a con	https://huggingfa	https://huggingfa	ind jav khm lao tgl min mya sun tha vie zlm ceb war cbk bcl	N/A	Language Model	Language	Multi-domain	Other (other)	Crawling Machine-generat	Automatic & Man						23251368 2058 1013181 217842 348453 1429 865575 1554 20960550 102411180 238151 263890 23687 2 1	documents	149598921	University of Ore Adobe Research	CulturaX: A Clear	https://arxiv.org/a	2023	Preprint	Free upon reque	mC4 and OSCAR	Yes	No	2023-11-16 15:24	Approved	culturax	https://github.com
76	CulturaY		CulturaY: A Large From the team th	https://huggingfa	https://huggingfa	mya fil zlm vie ind tha	N/A	Language Model	Language	Mixed	Creative Commo	Crawling Crowdsourced	Automatic						16330227 8473141 4888894 4493567 575221 513729	documents	35274779	Independent	CulturaY: A Large	https://huggingfa	2024	No publication	Free		No	No	2024-03-16 11:37	Approved	culturay	https://github.com
77	Customer Review (Natasha Skinca		This dataset is cc	https://drive.goog		ind		Emotion Classific	Language	Social media	Unknown (unkno	Crawling Expert-generatec				87,120	37,143	37,143		tweets	161406	Tweet @Natasha	CLASSIFICATIOI	https://jurnal.uns.	2017		Free		No	No		NusaCrowd carry-over	sentiment_natha	
78	CVSS	Indonesian	We introduce CV	https://github.com		ind		Speech-to-Speec	Language	General	Creative Commo	Crowdsourced Expert-translated				3	2	2		hours	6.3	IR-NLP Lab Fasil	CVSS Corpus an	https://arxiv.org/p	2022		Free	Common Voice,				NusaCrowd carry-over	cvss	
79	DaMuEL		DaMuEL is a larg	https://lindat.mff.c		ind vie	N/A	Entity Linking POS Tagging Morphological Inf Named Entiy Rec	Language	Wikipedia	Creative Commo	Crawling Machine-generat	None						273800000	tokens	273800000	Charles Universit	DaMuEL: A Large	https://arxiv.org/p	2023	Preprint	Free		No	No	2024-01-16 12:05	Approved	damuel	https://github.com
80	DeepLontar		DeepLontar is a c	https://doi.org/10.		ban	Bali, Indonesia	Optical Characte	Language Vision	Culture & heritage	Creative Commo	Expert-generatec	Manual (full)			520	N/A	80		images	600	Institut Teknologi Universitas Uday Universitas Hindu Institut Bisnis dan	DeepLontar data	https://www.natur	2022	Nature	Free	Lontar manuscrip	No	No	2023-12-23 17:47	Approved with notes	deep_lontar	https://github.com
81	Dengue Filipino		Benchmark datas	https://huggingfa	https://huggingfa	fil	Phillipines	Text Classificatio	Language	Social media	Unknown (unkno	Crowdsourced	Automatic & Man			4015	500	500		tweets	5015	De La Salle Univ	Intelligent Dengu	https://ieeexplore	2018	ICA	Free		No	No	2023-11-20 01:58	Approved with notes	dengue_filipino	https://github.com
82	Duolingo STAPLE 2020	aws_baseline gold	This dataset is pr	https://dataverse.		eng vie	N/A	Machine Translat Paraphrasing	Language	Conversational	Creative Commo	Crowdsourced Machine-translat	Automatic & Man			3500	500	500		sentences	4500	Duolingo	Simultaneous Tra	https://aclantholo	2020	ACL	Free upon reque		No	No	2024-03-07 08:03	Approved	duolingo_staple_	https://github.com

No	Dataset name	Subset(s)	Dataset description	Dataset URL	HuggingFace URL	Dataset language(s)	Dataset collection region	Dataset task(s)	Dataset modality	Dataset domain(s)	Dataset license	Dataset annotation collection style	Dataset annotation validation style	Number of annotators per sample	Inter-annotator agreement	Train split data size	Validation split data size	Test split data size	No split data size	Data size unit	Total data size	Dataset provider	Dataset paper title	Dataset paper URL	Dataset or dataset paper publish year	Publication venue	Dataset access	Derived from...	Does the data have PII (personally identifiable information)?	Does the data have sensitive content/information?	Submission Date	Approval Status	Dataloader name	GitHub Issue URL
83	eBible		Curated corpus c	https://github.con	https://huggingfa	kne ptu msk blz txq ind mkn tcz beu san tiy lbk smk mqj prf yva tam amk alp tdt tha cbk heg zyp mbb kkl tgl hlt blw yka atb hvn bkx mya eng ilo por sgb bgs kmk ceb cgc aaz atd tbl row xnn llg lex tet clu rgu agt vie cmn bps csy att cth kje cek msm zlm mbs obo msb mbt abx dgc nbq bkd nfa agn ebk bpr wrs	N/A	Language Model Machine Translat	Language	Religion	Creative Commo	Crawling Expert-generatec	None						N/A	sentences	0	University of Sou	N/A		2022	No publication	Free		No	No	2023-12-26 14:4	Approved	ebible	https://github.con
84	EmoT (IndoNLU Split)		EmoT is an emot	https://github.con	https://huggingfa	ind		Emotion Classific	Language	social media	Creative Commo	Expert-generatec				3,521	440	442		sentence pairs	4403	Tokyo Institute of	Emotion classific twitter dataset.	https://ieeexplore	2018		Free	Twitter	No	No		NusaCrowd carry-over	emot	
85	EmotCMT		EmotCMT is a In https://thesai.org	https://github.con		ind eng		Emotion Classific	Language	Social media	Unknown (unkno	Crawling Expert-generatec				582	0	0		sentences	582	FMIPA Universita	Normalisation of	https://thesai.org	2021		Free	Normalization of	No	No		NusaCrowd carry-over	emotcmt	
86	EMoTES-3K	Tagalog English	This dataset feat	https://huggingfa	https://huggingfa	tgl eng	Philippines Philippines	Commonsense R	Language	Ethics	Apache license 2	Expert-generatec	Manual (full)						2905 2905	examples	5810	Independent University of the	Emotion-based M	https://rootroo.co	2023	NLP4DH	Free		No	No	2023-12-22 15:0	Approved	emotes_3k	https://github.con
87	Emotion Indonesian Public Opinion		The dataset is fo	https://github.con		ind		Emotion Classific	Language	Social media	Creative Commo	Expert-generatec Expert-translated				7,080	0	0		tweets	7080	Ahmad Izzan	Emotion dataset	https://www.scien	2022		Free	Twitter	No	No		NusaCrowd carry-over	emotion_id_opini	
88	ETOS		ETOS (Ejaan oT	https://github.con		ind	Indonesia	POS Tagging	Language	General	GNU Affero Gen	Crawling Expert-generatec	Manual (full)						200	sentences	200	University of Indc	A Comparison of	https://ieeexplore	2022	ISRITI	Free		No	No	2024-01-08 18:1	Approved	etos	https://github.con
89	FacQA		The goal of the F	https://github.con	https://huggingfa	ind		Question Answer	Language	news articles	Creative Commo	Expert-generatec				2,495	311	311		documents	3117	Louis Owen	A Machine Learn Indonesian Ques	https://www.resea	2007		Free	Wikipedia	No	No		NusaCrowd carry-over	facqa	
90	Fake News Filipino		Fake news article	https://github.con	https://huggingfa	fil	Phillipines	Hoax Detection	Language	News	Unknown (unkno	Expert-generatec	Manual (full)			2244	0	962		documents	3206	De La Salle Univ	Localization of Fa	https://aclantholo	2019	LREC	Free		No	No	2023-11-20 01:5	Approved	fakenews_ph	https://github.con
91	Filipino Age-of-Acquisition Words		The dataset cont	https://osf.io/j42g		fil	Philippines	Machine Translat	Language	General	Creative Commo	Expert-generatec Machine-generat	Automatic						N/A	tokens	0	University of Oxfo University of the	TalkTogether Age	https://osf.io/gnjn	2021	Technical Report	Free upon reque		No	No	2023-11-16 22:4	Approved	filipino_words_ac	https://github.con
92	Filipino Gay Language (Swardspea		The dataset cont	https://github.con		fil	Philippines	Language Model	Language	General Mixed	Creative Commo	Crawling	None						N/A	tweets	0	National Universi	Witchebelles Ana	https://national-u	2016	NU Journal of Sc	Free upon reque		No	No	2023-11-15 06:0	Approved	filipino_gay_lang	https://github.con
93	Filipino Hate Speech Tiktok - Text		The dataset cont Labeling: 0 - no h	https://github.con		tgl eng	Philippines	Hate Speech Det	Language	Election Politics	Creative Commo	Crawling	Manual (full)	3	0.75	10038	4246	4254		comments	18538	National Universi	A BERT-based H	https://dl.acm.org	2021	ICSET	Free upon reque		No	Yes	2023-11-15 05:5	Approved	filipino_hatespee	https://github.con
94	Filipino Product Reviews (Lazada)		The dataset cont	https://github.con		fil	Philippines	Sentiment Analys	Language	Reviews Social media	Creative Commo	Crawling	Automatic						N/A	paragraphs	0	Independent	No paper associa	https://github.con	2022	No publication	Free		No	No	2023-11-19 09:0	Approved	lazada_review_fi	https://github.con
95	Filipino Slang Spelling Normalizatic		This dataset cont	https://github.con		fil	Philippines	Lexical Normaliz	Language	Conversational Social media	Unknown (unkno	Expert-generatec	Manual (full)	3	0.838	17965	993	1007		examples	19965	Yale University	Look Ma, Only 4	https://aclantholo	2022	EMNLP	Free		No	No	2023-11-01 10:5	Approved	filipino_slang_no	https://github.con
96	FilWordNet		Filipino WordNet	https://drive.goog		fil	Philippines	Wordnet	Language	wordnet	Unknown (unkno	Crowdsourced	Manual (full)						15929	instances	15929	De La Salle Univ	Introducing Filipi	https://www.resea	2010	Global Wordnet (	Free upon reque	Princeton WordN	No	No	2024-01-05 13:5	Approved	filwordnet	https://github.con
97	Fleurs	Fleurs Burmese Fleurs Cebuano Fleurs Filipino Fleurs Indonesian Fleurs Javanese Fleurs Khmer Fleurs Lao Fleurs Malay Fleurs Thai Fleurs Vietnamese	Fleurs dataset is	https://huggingfa	https://huggingfa	mya ceb fil ind jav khm lao zlm tha vie	Multiple countries	Automatic Speec Spoken Languag	Speech	Wikipedia	Creative Commo	Crowdsourced	None			1,509 1,509 1,509 1,509 1,509 1,509 1,509 1,509 1,509 1,509 1,509 1,509 1,509	150 150 150 150 150 150 150 150 150 150 150 150 150	350 350 350 350 350 350 350 350 350 350 350 350 350		sentences	20090	Google	XTREME-S: Eval	https://www.isca-	2022	Interspeech	Free	Flores-200	No	No	2023-11-07 04:1	Approved	fleurs	https://github.con
98	FLORES-200	ace ban bjn bug ceb ilo ind jav kac khm lao lus min mya pag shn sun tgl tha vie war zsm	The creation of F	https://github.con		ace ban bjn bug ceb ilo ind jav kac khm lao lus min mya pag shn sun tgl tha vie war zsm	N/A	Machine Translat	Language	News articles Wikipedia Wikidata	Creative Commo	Crawling Expert-generatec	Automatic & Man	2	1	N/A	997	1012		sentence pairs	2009	Meta	The FLORES-10	https://arxiv.org/a	2022	TACL	Free		Yes	No	2023-11-12 17:3	Approved	flores200	https://github.con
99	FSL-105		FSL-105 is a vide	https://data.menc		psp	nal Capital Region	Video-to-Text Re Video Captioning	Vision	Conversational	Creative Commo	Expert-generatec	Manual (full)			68	27	55		examples	150	De La Salle Univ	Recognizing Filip	https://ieeexplore	2022	Institute of Electr	Free		No	No	2023-11-01 08:1	Approved	fsl_105	https://github.con
100	FTC-200 (SIB-200)	ace ban bjn bug ceb ilo ind jav kac khm lao lus min mya pag shn sun tgl tha vie war zsm	SIB-200 is a topi	https://github.con	https://huggingfa	ace ban bjn bug ceb ilo ind jav kac khm lao lus min mya pag shn sun tgl tha vie war zsm	N/A	Text Classificatio	Language	Health Politics Sports Science travel Uncategorized Entertainment	Apache license 2	Expert-generatec Machine-generat	Manual (partial)	4	0.44	701 701 701 701 701 701 701 701 701 701 701 701 701 701 701 701 701 701 701 701 701 701	99 99 99 99 99 99 99 99 99 99 99 99 99 99 99 99 99 99 99 99 99 99	204 204 204 204 204 204 204 204 204 204 204 204 204 204 204 204 204 204 204 204 204 204	485936	examples	22088	Amazon Masakhane University Colleg University of Torc Saarland Univers University of Onta	SIB-200: A Simpl	https://arxiv.org/p	2023	Preprint	Free	Flores-200	No	No	2023-12-27 04:5	Approved	sib200	https://github.con
101	GATITOS	ace ban bbc bew bjn bts btx bug cnh hil iba ilo kac lus mad mak meo min pag pam shn tet war	The GATITOS (G	https://github.con		ace ban bbc bew bjn bts btx bug cnh hil iba ilo kac lus mad mak meo min pag pam shn tet war	N/A	Machine Translat	Language	Mixed	Creative Commo	Expert-generatec	None						4000	sentence pairs	4000	Google	Bilex Rx: Lexical	https://arxiv.org/a	2023	EMNLP	Free		No	No	2023-11-28 15:4	Approved	gatitos	https://github.con
102	generated_reviews_enth		This dataset (refe	https://github.con		tha eng	N/A	Machine Translat	Language	Reviews	Creative Commo	Machine-generat	Manual (full)			42073	4772	5245		sentence pairs	52090	Vistec-AI	SCB-MT-EN-TH-	https://arxiv.org/p	2020	Preprint	Free		No	No	2023-11-17 14:2	Approved with notes	generated_review	https://github.con
103	GKLMIP Khmer News Dataset		The GKLMIP Khr	https://github.con		khm	N/A	Topic Classificati	Language	News	Unknown (unkno	Crawling	None			4296	1434	1436		documents	7166	Guangdong Univ	Pretrained Model	https://ieeexplore	2021	Tsinghua Science	Free		No	No	2024-01-20 00:1	Approved	gklmip_newsclas	https://github.con
104	GKLMIP Product Sentiment		The GKLMIP Pro	https://github.con		mya	China	Sentiment Analys	Language	E-commerce	Unknown (unkno	Crawling	None			2492	358	716		sentences	3566	Guangdong Univ	Pre-trained Mode	https://link.spring	2021	International Con	Free		No	No	2024-01-17 00:3	Approved	gklmip_sentimen	https://github.con
105	GlobalVoices		A parallel corpus	https://opus.nlpl.e		fil ind khm mya	N/A	Machine Translat	Language	News articles	Creative Commo	Expert-generatec	None						(Large number o	sentence pairs	0	Uppsala Universi Nordic Language	Parallel Data, To	http://www.lrec-c	2012	LREC	Free		No	Yes	2023-12-19 18:3	Approved with notes	globalvoices	https://github.con
106	GlobalWoZ	E&F_id E&F_th E&F_vi F&E_id F&E_th F&E_vi F&F_id F&F_th F&F_vi	This is the data o the target langua	https://entuedu-n		ind tha vie	N/A	Dialogue System	Language	Hotel reviews Restaurant Attraction Hospital Police Taxi Train	Unknown (unkno	Crawling Machine-generat	Manual (partial)						10437 10437 10437 10437 10437 10437 10437 10437 10437	dialogues	93933	Alibaba Group Nanyang Techno	GlobalWoZ: Glob	https://aclantholo	2022	ACL	Free	MultiWoZ 2.2	Yes	No	2023-11-16 15:0	Approved	globalwoz	https://github.con
107	Glot500-c		Glot500-c is a co	https://github.con	https://huggingfa	bsb iba ind eng zsm khm lao tha tdt por ace fil vie tih mya bcl ceb zlm jav sun bjn min tgl tam hil ilo kac war ahk dtp ksw lhu pag cmn pam bbc ban sxn nia btx gor mad bts mbb prk ibg bhw ifb ifa mrw	N/A	Language Model Named Entity Re POS Tagging Text Classificatio	Language	Multi-domain	Other (other)	Crawling Crowdsourced Expert-generatec Machine-generat Expert-translated Machine-translat	Automatic & Man						1500000000	sentences	1500000000	LMU München	Glot500: Scaling	https://aclantholo	2023	ACL	Free		No	No	2024-03-22 16:5	Approved	gloot500_c	https://github.con
108	GlotStoryBook		The GlotStoryBo	https://github.con	https://huggingfa	khg khm mya por tet tha eng	N/A	Language Model	Language	Books	Other (other)	Crawling	Automatic						N/A	instances	0	Munich Center fo	GlotLID: Langua	https://arxiv.org/a	2023	EMNLP	Free	Global Storybook	No	No	2023-11-16 13:0	Approved	glotstorybook	https://github.con
109	GNOME		A parallel corpus	https://opus.nlpl.e		eng vie mya ind tha tgl zlm lao	N/A	Machine Translat	Language	Technology	Unknown (unkno	Expert-generatec	Manual (full)						N/A	sentence pairs	0	Uppsala Universi GNOME Translat	Parallel Data, To	https://aclantholo	2012	LREC	Free		No	No	2023-12-31 16:0	Approved with notes	gnome	https://github.con
110	Gowajee Corpus		The Gowajee cor	https://github.con		tha	Thailand	Automatic Speec	Language Speech	Smart home	MIT (mit)	Crowdsourced Expert-generatec	Manual (full)						20308	utterances	40616	Chulalongkorn U	Gowajee Corpus	https://github.con	2023	Technical Report	Free		No	No	2024-03-20 22:1	Approved	gowajee	https://github.con

No	Dataset name	Subset(s)	Dataset description	Dataset URL	HuggingFace URL	Dataset language(s)	Dataset collection region	Dataset task(s)	Dataset modality	Dataset domain(s)	Dataset license	Dataset annotation collection style	Dataset annotation validation style	Number of annotators per sample	Inter-annotator agreement	Train split data size	Validation split data size	Test split data size	No split data size	Data size unit	Total data size	Dataset provider	Dataset paper title	Dataset paper URL	Dataset or dataset paper publish year	Publication venue	Dataset access	Derived from...	Does the data have PII (personally identifiable information)?	Does the data have sensitive content/information?	Submission Date	Approval Status	Dataloader name	GitHub Issue URL
111	Grambank		Grambank is a gl	https://grambank		mrh pac ugo ium ctd ind mrz tyn blf rth jeh blz slu dsn gay wtw cmn sas nbe agt ify mnb zrs pgu auu twe dtb nqm kxm brv wsa kvv mnw sbl dnw cua tlb mpz mbs suc smk tft wbw ekg yva plh lcp kns duo lbt xnn bve cfm gei sfe krj yka eng msb lmy mnx kmk skh mtg bzu cjm stb twu nbq pmf lhi lrt min mya zsm ptu lbo pez bgs bdg mej duu bth wrp msi tgl pce slp dms mbi ahk plc msk atb pnx nir pag iba mbf rad msm ahh mwt wlo jav bgr kne mkm tbw adr kvo sob hvn spu bvz cja rbb kgq kys mra lao cmo kml bkx szc bew mso weo txx szp gbi nlc adn xmz	N/A	Wordnet	Language	General Linguistic Features	Creative Commo	Expert-generated	None						195	features	195	Max Planck Instit Max Planck Institu	Grambank reveal	https://www.scien	2023	Science Advance	Free		No	No	2024-03-31 09:2	Approved	grambank	https://github.com
112	Graves' Disease Chatbot Dataset i	Indonesian	A dataset consist	https://dataverse.		ind		Chatbot	Language	Health	Creative Commo	Expert-generated				607	0	0		sentences	607	Paolo Pareti			2022		Free		No	No		NusaCrowd carry-over	grave_disease_c	
113	Hate Speech Filipino Tweet		The dataset usec Data preprocessi	https://huggingfa	https://huggingfa	fil	N/A	Hate Speech Det	Language	Social media Election	Unknown (unkno	Crowdsourced	Manual (full)	4	0.86	10000	4230	10000		tweets	24230	De La Salle Univ	Hate Speech in F	https://www.resea	2019	Philippine Compu	Free		No	No	2024-02-02 17:2	Approved	filipino_hatespee	https://github.com
114	HoASA (IndoNLU Split)		HoASA is an asp	https://github.com	https://huggingfa	ind		Aspect Based Se	Language	hotel reviews	Creative Commo	Crawling Expert-generated				7,560	0	1,890		sentences	9450	Google	Multi-label Aspec	https://ieeexplore	2019		Free	Online Platform	No	No		NusaCrowd carry-over	hoasa	
115	HPLTDatasets v1.2		The dataset is pa monolingual and and builds efficie	https://hplt-projec	https://huggingfa	ind zlm tha mya fil vie	N/A	Language Model	Language	General Web	Creative Commo	Crawling	Automatic						31.42M 469.98K 8.19M 239.47K 585.24K	documents	0	University of Edir University of Hels Charles Universit University of Oslo University of Turk Prompsit Langua CESNET Sigma2 HPC cen	HPLT: High Perfo	https://aclantholo	2023	European Associ	Free		No	No	2024-03-10 00:0	Approved	hplt	https://github.com
116	HSE Thai Corpus		he HSE Thai Cor The data for the (This dataset cont	https://www.kagg		tha	Thailand	Language Model Language Identif	Language	General Mixed Multi-domain News articles	Apache license 2	Crawling	None						N/A	tokens	0	HSE University	No paper	https://www.kagg	2017	No publication	Free		No	No	2023-11-21 06:5	Approved	hse_thai	https://github.com
117	Human Instructions - Indonesian (w		Human Instructio	https://www.kagg		ind		Knowledge Base	Language	General	Creative Commo	Crawling Expert-generated				39,246	0	0		documents	39246	Google	Vocabulary Align	https://paolopare	2017		Free		No	No		NusaCrowd carry-over	id_wikihow	
118	iapp_wiki_qa_squad		iapp_wiki_qa_sq	https://github.com	https://huggingfa	tha	Thailand	Question Answer	Language	Wikipedia	MIT (mit)	Crawling	None			5761	742	739		instances	7242	iAPP Technology	-	https://zenodo.or	2021	No publication	Free		Yes	No	2023-12-13 22:3	Approved	iapp_squad	https://github.com
119	ICON		ICON (Indonesia	https://github.com		ind	Indonesia	Constituency Par	Language	News articles Wikipedia	Creative Commo	Expert-generated	Manual (partial)			8000	1000	1000		sentences	10000	Bandung Institute AI Singapore Prosa.ai	ICON: Building a	https://aclantholo	2023	TLT	Free		No	No	2023-12-01 16:0	Approved	icon	https://github.com
120	ID Abusive		ID Abusive is a T	https://github.com	-	ind		Hate Speech Det	Language	social media	Creative Commo	Crawling Expert-generated				2,016	0	0		tweets	2016	Universitas Daru	A Dataset and Pr	https://www.scien	2018		Free	Twitter				NusaCrowd carry-over	id_abusive	
121	ID Abusive Online News Comment		The dataset cons	https://github.com		ind		Abusive Languag	Language	Commentary	Unknown (unkno	Crawling				3,184	0	0		comments	3184	Wilson Wongso &	Abusive Languag	https://ieeexplore	2019		Free					NusaCrowd carry-over	id_abusive_news	
122	ID Coreference Resolution		ID Coreference r	https://github.com		ind		Coreference Res	Language	News articles	Unknown (unkno	Crawling Expert-generated				759	0	108		sentences	867	Institut Teknologi	Experiments on c	https://ieeexplore	2017		Free	IDENTIC	No	No		NusaCrowd carry-over	id_coref	
123	ID Multilabel HS		ID Multilabel HS	https://github.com	-	ind		Hate Speech Det	Language	social media	Creative Commo	Crawling Expert-generated				13,169	0	0		tweets	13169	Institut Teknologi	Multi-label Hate S	https://aclantholo	2019		Free	ID Abusive, Twitt	No	No		NusaCrowd carry-over	id_multilabel_hs	
124	ID Newspapers 2018		This dataset aims	https://github.com	https://huggingfa	ind	N/A	Language Model	Language Vision	News	Creative Commo	Crawling Expert-generated	None						60GB (original, i 2.2GB (text only) 1GB (text only, c	MB	0	Independent	N/A		2018	No publication	Free		No	No	2024-01-29 12:4	Approved	id_newspaper_2	https://github.com
125	ID Quora Question Pairs		Quora Question I	https://github.com		ind		Paraphrasing	Language	General	Other (other)	Machine-translat				134,084	14,927	0		sentence pairs	149011	Multiple Insitution			2021		Free	https://quoradata				NusaCrowd carry-over	id_qqp	
126	ID Short Answer Grading		ID Short Answer	https://github.com		ind		Short Answer Gra	Language	Education	Unknown (unkno	Expert-generated				7,605	0	1,560		sentences	9165	Multiple Institutio	Short Answer Gra	https://inacl.id/jou	2020		Free		No	No		NusaCrowd carry-over	id_short_answer_	
127	id-en-code-mixed		This dataset cont	https://github.com		ind eng		Token-level langu	Language	Social media	Creative Commo	Crawling Expert-generated				22,736	0	0		tokens	22736	KAIST	Normalization of	https://github.com	2019		Free	Twitter	No	No		NusaCrowd carry-over	id_en_code_mixe	
128	ID-HSD-Nofaaulia	Indonesian	There have been	https://github.com	https://github.com	ind		Hate Speech Det	Language	Social media	Unknown (unkno	Crawling Expert-generated				815	0	91		documents	906	BINUS	Hate Speech Det	https://dl.acm.org	2022		Free	Facebook	No	No		NusaCrowd carry-over	id_hsd_nofaaulia	
129	ID-HSD-Riomulia		ID-HSD-RioMulia 2017, whose sele	https://github.com	-	ind		Hate Speech Det	Language	social media	Unknown (unkno	Crawling Expert-generated				713	0	0		tweets	713	Universitas Al Az	Hate Speech Det	https://ieeexplore	2017		Free	Twitter	No	No		NusaCrowd carry-over	id_hatespeech	
130	id-vaccines-tweets		Dataset containir	https://github.com		ind	N/A	Sentiment Analys	Language	Commentary Health Reviews Social media aster Preparedne Hospital	Unknown (unkno	Crawling	Manual (full)	1	1	4183				sentences	4183	Independent Bandung Institute	Analisis Sentimer	https://github.com	2021	Preprint	Free	Twitter (now X)	Yes	Yes	2023-12-11 11:3	Approved	id_vaccines_twe	https://github.com
131	Identic (Larasati, 2012)		IDENTIC is an In	https://lindat.mff.	-	eng ind		Machine Translat	Language	multi domain	Unknown (unkno	Crawling Expert-generated				45,000	0	0		sentences	45000	Bangladesh Univ	IDENTIC Corpus	https://aclantholo	2012		Free	-				NusaCrowd carry-over	identic	
132	identifikasi-bahasa		The identifikasi-b	https://github.com		ind jav abs	N/A	Language Identif	Language	Conversational Social media	Apache license 2	Crawling Expert-generated	Manual (full)			39511	N/A	16933		sentences	56444	Independent	PERBANDINGAN	https://ejournal.u	2021	JIKO (Jurnal Info	Free		No	No	2024-01-20 09:2	Approved	identifikasi_baha	https://github.com
133	IDK-MRC		IDK-MRC is an Ir (Note: the paper	https://anonymou		ind		Question Answer	Language	General Wikipedia	Creative Commo	Machine-generat Expert-generated				9,332	764	844		paragraph, quest	10940	University of Can	IDK-MRC: Unans	https://arxiv.org/a	2022		Free	IDK-MRC datase	No	No		NusaCrowd carry-over	idk_mrc	
134	IDK-MRC-NLI	Indonesian	The IDK-MRC-NI	https://huggingfa	https://huggingfa	ind	N/A	Natural Languag	Language	General Wikipedia	Creative Commo	Machine-generat	Manual (partial)			18665	1529	1689		sentence pairs	21883	University of Indo Universitas Indor	Verification of QA		2024	NAACL	Free	IDK-MRC	No	No	2024-03-21 23:0	Approved	idk_mrc_nli	https://github.com
135	Idn-tagged-corpus-CSUI		Id-tagged-corpus	https://github.com	https://huggingfa	ind		POS Tagging	Language	news articles	Creative Commo	Crawling Expert-generated				8,000	1,000	1,029		sentences	10029	University of Utal	Designing an Ind	https://ieeexplore	2014		Free	-	No	No		NusaCrowd carry-over	idn_tagged_corp	
136	idner-news-2k		A dataset of Indo	https://github.com		ind	N/A	Named Entity Re	Language	News articles	MIT (mit)	Crowdsourced	Manual (full)			1464	367	509		documents	2340	Tokyo Metropolita	Towards a Stand	https://aclantholo	2020	AACL SRW	Free	https://github.com	No	No	2024-03-15 19:5	Approved	idner_news_2k	https://github.com
137	IJELID		This is a code-mi MIX_ID_EN (mix	https://github.com		ind jav eng	Indonesia Java	Token-level langu	Language	General Mixed Multi-domain Social media	Creative Commo	Crawling	Automatic & Man	2	0.9964	281403	35175	35177		tweets	351755	Universiti Brunei Universitas Islam	Corpus creation a	https://peerj.com/	2023	PeerJ Computer	Free		No	No	2023-10-31 21:5	Approved	ijelid	https://github.com
138	IMDb Javanese		Large Movie Rev	https://huggingfa	https://huggingfa	jav		Sentiment Analys	Language	Reviews	Unknown (unkno	Machine-translat				25,000	0	25,000		sentences	50000	Telkom University	Causal and Mask	https://ieeexplore	2021		Free	IMDb Dataset	No	No		NusaCrowd carry-over	imdb_jv	
139	Indo Wiki Paralel Corpora	Javanese Sundanese Minangkabau	Manually aligned	https://github.com		ind sun jav min		Machine Translat	Language	Wikipedia	Unknown (unkno	Crawling Expert-generated				2,422	0	0		sentence pairs	2422	Telkom University	Creating Indones	https://ieeexplore	2014		Free	Wikipedia	No	No		NusaCrowd carry-over	id_wiki_parallel	
140	Indo_MultiModal_CC_12M	Indonesian	Conceptual 12M	https://huggingfa	https://huggingfa	ind		Image-to-Text Ge Text-to-Image Ge VL Modeling	Vision Language	General	Other (other)	Machine-translat				1	0	0		GB	1.3	Clova AI, NAVER	Conceptual 12M:	https://arxiv.org/a	2022		Free	Conceptual 12M				NusaCrowd carry-over	id_mm_cc_12m	
141	Indo_MultiModal_LAION	Indonesian	Indo_MultiModal_	https://huggingfa	https://huggingfa	ind		Image-to-Text Ge Text-to-Image Ge VL Modeling	Vision Language	General	Creative Commo	Machine-translat				7	0	0		GB	6.7	Facebook AI	LAION-400M: Op	https://arxiv.org/a	2022		Free	LAION-400	No	No		NusaCrowd carry-over	id_mm_laion	
142	Indo_MultiModal_PMD_ID	Indonesian	Introduced in the	https://huggingfa	https://huggingfa	ind		Image-to-Text Ge Text-to-Image Ge VL Modeling	Language Vision	General	Other (other)	Machine-translat								GB	0	Columbia Univer	Flava: A foundati	https://arxiv.org/a	2022		Free	Conceptual Capt	No	No		NusaCrowd carry-over	id_mm_pmd	
143	indo_story_cloze		This is the first c	https://huggingfa	https://huggingfa	ind	Indonesia	Commonsense R	Language	General	MIT (mit)	Expert-generated	Manual (partial)			1000	200	1135		documents	2335	The University O	Cloze Evaluation	https://aclantholo	2022	ACL Workshop	Free	NA	No	No	2023-11-14 01:5	Approved	indo_story_cloze	https://github.com
144	indo-law		This dataset con	https://github.com		ind		Legal Classificati	Language	Law	Unknown (unkno	Crawling Expert-generated				22,630	0	0		documents	22630	IT University of C	Predicting the Ca	https://www.mdpi	2022		Free		No	No		NusaCrowd carry-over	indo_law	
145	Indo4B		Indo4B is an Indo	https://storage.gc	-	ind		Language Model	Language	multi domain	Creative Commo	Crawling				4	0	0		tokens	3.6	Google Research	IndoNLU: Bench	https://aclantholo	2020		Free	Multiple Online S	No	No		NusaCrowd carry-over	indo4b	
146	Indo4B Plus		Indo4BPlus is an	https://storage.gc	-	ind sun jav		Language Model	Language	multi domain	Creative Commo	Crawling				4	0	0		tokens	4	Language Techn	IndoNLG: Bench	https://aclantholo	2021		Free	Multiple Online S	No	No		NusaCrowd carry-over	indo4b_plus	
147	IndoAbusive		IndoAbusive cont	https://github.com		ind	Indonesia	Abusive Languag	Language	Mixed	Creative Commo	Crawling	Manual (full)						5461	tweets	5461	Universitas Islam	Identifikasi Konte	http://inacl.id/jour	2019	Jurnal Linguistik	Free		No	No	2024-05-12 23:3	Approved	indoabusive	https://github.com
148	IndoACD	an Adult Content I	This dataset cont	https://github.com		ind	Indonesia	Text Classificatio	Language	Tweet	Creative Commo	Crawling	Automatic & Man			38037	3260	13041		tweets	54338	Universitas Islam	Adult Content De	https://ieeexplore	2023	IEEE	Free		No	Yes	2024-05-12 23:5	Approved	indoacd	https://github.com
149	IndoAMR		IndoAMR is anno	https://github.com		ind		Abstract Meaning	Language	News articles	Unknown (unkno	Crawling Expert-generated				700	0	300		sentences	1000	Kata.ai	Parsing Indonesi	https://arxiv.org/p	2021		Free		No	No		NusaCrowd carry-over	indoamr	
150	IndoCamRest		IndoCamRest is	https://github.com		ind	Indonesia	Dialogue System	Language	Restaurant	Creative Commo	Crowdsourced	Manual (partial)						676	dialogues	676	Bandung Institute	IndoToD: A Multi-	https://arxiv.org/p	2023	SEALP	Free	CamRest676	Yes	No	2023-11-14 17:3	Approved	indocamrest	https://github.com
151	IndoCollex		List of Colloquial	https://github.com		ind		Morphological Inf	Language	General	MIT (mit)	Expert-generated Expert-translated				1,637	182	193		sentence pairs	2012	TUFS	IndoCollex: A Tes	https://aclantholo	2021		Free		No	No		NusaCrowd carry-over	indocollex	
152	IndoCoref		IndoCoref is a co	https://github.com	-	ind		Coreference Res	Language	general	MIT (mit)	Expert-generated				0	0	201		documents	201	TUFS	A Multi-Pass Siev	https://aclantholo	2021		Free	Wikipedia	No	No		NusaCrowd carry-over	indocoref	
153	IndoCulture		IndoCulture is air	https://huggingfa	https://huggingfa	ind	Indonesia	Commonsense R	Language	General Religion Culture & heritage	Creative Commo	Expert-generated	Manual (full)	150					2429	instances	2429	The University O MBZUAI University of Indo	IndoCulture: Expl	https://arxiv.org/p	2024	Preprint	Free	NA	No	No	2024-04-03 12:3	Approved	indoculture	https://github.com
154	IndoLEM NTP		IndoLEM next tw	https://github.com	-	ind		Next Tweet Predi	Language	social media	Creative Commo	Crawling Expert-generated				5,681	811	1,890		instances	8382	David Moeljadi	IndoLEM and Ind	https://aclantholo	2020		Free	Twitter	No	No		NusaCrowd carry-over	indolem_ntp	
155	IndoLEM Sentiment Analysis		IndoLEM Sentim	https://github.com	-	ind		Sentiment Analys	Language	social media	Creative Commo	Crawling Expert-generated				3,638	399	1,011		sentences	5048	David Moeljadi	IndoLEM and Ind	https://aclantholo	2020		Free	Twitter	No	No		NusaCrowd carry-over	indolem_sentime	
156	IndoLEM Tweet Ordering		IndoLEM tweet o	https://github.com	-	ind		Tweet Ordering	Language	social media	Creative Commo	Crawling Expert-generated				5,327	760	1,521		instances	7608	David Moeljadi	IndoLEM and Ind	https://aclantholo	2020		Free	Twitter				NusaCrowd carry-over	indolem_tweet_o	
157	Indoler		The dataset cont	https://github.com		ind	Indonesia	Named Entity Re	Language	Legal	Unknown (unkno	Crawling Machine-generat	Automatic			695	30	268		examples	993	University of Indo	Legal Entity Reco	https://ieeexplore	2020	ICACSIS	Free		Yes	No	2024-01-04 02:2	Approved	indoler	https://github.com
158	IndoMMLU		IndoMMLU is the	https://huggingfa	https://huggingfa	ind ban mad mak sun jav bjn abl nij	Indonesia	Question Answer	Language	Education	Creative Commo	Crawling	Manual (partial)					11478 471 295 186 1157 992 144 147 109	14,979	examples		MBZUAI	Large Language	https://arxiv.org/p	2023	EMNLP	Free		No	No	2023-11-14 02:0	Approved	indommlu	https://github.com
159	IndoModal		IndoModal is a d	https://github.com		ind	Indonesia	Word Sense Disa	Language	Mixed News articles Wikipedia	Creative Commo	Expert-generated	None	3					600	sentences	600	Tokyo University	Annotation of mo	https://www.anlp.	2024	Annual Meeting c	Free	Leipzig Corpora (	No	No	2024-03-22 16:1	Approved	indomodal	https://github.com
160	IndoNER-Tourism		This dataset is de - B-WIS: Beginni - I-WIS: Continua - B-LOC: Beginni - I-LOC: Continua - B-FAS: Beginni - I-FAS: Continua - O: Non-entity o	https://github.com		ind	Indonesia	Named Entity Re	Language	Tourism	Academic Free L	Crawling	Manual (full)	2	0.996				116564	tokens	116564	Universitas Islam	Named Entity Re	https://www.inacl.	2023	Jurnal Linguistik	Free	https://github.com	No	No	2024-01-03 16:2	Approved	indoner_tourism	https://github.com
161	Indonesia BioNER		This dataset take Dataset contains	https://huggingfa	https://huggingfa	ind	Indonesia	Named Entity Re	Language	Health	BSD 3-clause Cl	Crawling	Manual (full)			1950	260	390		examples	2600	Institut Teknologi Universitas Budi	Pengenalan Entit	https://jtiik.ub.ac.i	2023	Jurnal Teknologi	Free		No	No	2024-02-02 16:2	Approved	bioner_id	https://github.com
162	Indonesia Sentiment Analysis Data		This dataset con See the dataset a	https://github.com		ind	Indonesia	Sentiment Analys	Language	Commentary Mixed Social media	Unknown (unkno	Crawling	None						10806	tweets	10806	Gadjah Mada Un	N/A	https://ridi.staff.u	2019	No publication	Free		No	No	2023-11-14 14:5	Approved	id_sentiment_an	https://github.com
163	Indonesia-Chinese-MTRobustEval		The dataset is cu	https://github.com		ind cmn	Indonesia	Machine Translat	Language	Social media	MIT (mit)	Crawling	Manual (full)			N/A	N/A	1001		sentence pairs	2002	Tianjin University	-		2023	Preprint	Free		Yes	Yes	2024-01-07 23:3	Approved	indonesia_chines	https://github.com
164	Indonesian Clickbait (CLICK-ID)		The CLICK-ID da	https://data.menc	https://huggingfa	ind		Clickbait Detectic	Language	news articles	Creative Commo	Crawling Expert-generated				15,000	0	0		headlines	15000	Institut Teknologi	CLICK-ID: A nove	https://www.scien	2020		Free	Online news pub	No	No		NusaCrowd carry-over	id_clickbait	
165	Indonesian Frog Storytelling Corpu		Indonesian writte	https://github.com		ind		Language Model	Language	Books	Unknown (unkno	Unknown				0		0		documents	0	FAISAL RAHUT	Usage of Indones	http://www.tufs.a	2014		Free		No	No		NusaCrowd carry-over	id_frog_story	
166	Indonesian GEC framework		A large corpus of	https://github.com		ind	Indonesia	Grammatical Erro	Language	Mixed	Unknown (unkno	Crawling Machine-generat	Automatic			5999634	424431	424420		examples	6848485	South China Univ Pamulang Univer	Automatic Correc	https://www.mdpi	2022	Applied Sciences	Free	CC100-Indonesia	No	No	2024-04-01 12:0	Approved	id_gec	https://github.com
167	Indonesian general domain MT En-		For the general d The Tatoeba data	https://github.com	-	eng ind		Machine Translat	Language	general	Creative Commo	Expert-generated				1,729,472	2,000	2,000		sentences	1733472	Universitas Indon	Benchmarking M	https://aclantholo	2020		Free	Wiki[x], Tatoeba,				NusaCrowd carry-over	indo_general_mt	
168	Indonesian Google Play Review		Indonesian Goog	https://huggingfa	https://huggingfa	ind		Sentiment Analys	Language	App review	Creative Commo	Crawling				7,028	3,012	0		sentences	10040	Jakarta AI Resea			2022		Free	-	No	No		NusaCrowd carry-over	id_google_play_r	
169	Indonesian Hoax News Detection		Indonesian Hoax	https://data.menc		ind		Hoax Detection	Language	News articles	Creative Commo	Expert-generated				600	0	0		documents	600	Jakarta AI Resea	Study of hoax ne	https://ieeexplore	2018		Free		No	No		NusaCrowd carry-over	id_hoax_news	
170	Indonesian News Corpus		This corpus cont	https://data.menc		ind	Indonesia	Topic Modeling Topic Classificati Sentiment Analys	Language	News articles	Creative Commo	Crawling	None						N/A	documents	0	Politeknik Negeri	Korpus berita dar	http://proceeding	2018	Prosiding Sentrin	Free		No	No	2024-04-02 16:4	Approved		
171	Indonesian News Dataset		An imbalanced d	https://github.com		ind	N/A	Text Classificatio Topic Classificati	Language	News Sports Business Technology Vehicle	Unknown (unkno	Crawling Expert-generated	None			6127	0	2627		instances	8754	Independent	N/A		2019	No publication	Free		No	No	2024-01-29 12:2	Approved	indonesian_news	https://github.com
172	Indonesian Poem Tweets		Indonesian Poem	https://huggingfa	https://huggingfa	ind		Language Model	Language	Social media	Creative Commo	Crawling				16,427	0	0		tweets	16427	Samsul Rahmadi			2022		Free	-				NusaCrowd carry-over	id_poem_tweets	
173	Indonesian PRONER	matically Labeled nually Labeled Da	A corpus for Indo	https://github.com		ind	Indonesia	Named Entity Re	Language	Product	Creative Commo	Crawling	Automatic & Man	1					2275 16853	sentences	19128	Telkom University	Corpus Developr	http://www.winlp.	2020	Data Science and	Free		No	No	2024-01-04 02:0	Approved	ind_proner	https://github.com
174	Indonesian religious domain MT En		Religious domain The Tanzil datase relatively-imbalar sentence in this c in this dataset wo as long as the otl the dataset in the remaining datase Another interestin is the localized na to Maryam, Gabr names are usuall We also find quit JW300 are missi the end sentence Lastly, we also fir "shalat", or repen	https://github.com	-	eng ind		Machine Translat	Language	religion	Creative Commo	Expert-generated				579,544	5,000	4,823		sentences	589367	Samsul Rahmadi	Benchmarking M	https://aclantholo	2020		Free	Tanzil, JW300, B				NusaCrowd carry-over	indo_religious_m	

No	Dataset name	Subset(s)	Dataset description	Dataset URL	HuggingFace URL	Dataset language(s)	Dataset collection region	Dataset task(s)	Dataset modality	Dataset domain(s)	Dataset license	Dataset annotation collection style	Dataset annotation validation style	Number of annotators per sample	Inter-annotator agreement	Train split data size	Validation split data size	Test split data size	No split data size	Data size unit	Total data size	Dataset provider	Dataset paper title	Dataset paper URL	Dataset or dataset paper publish year	Publication venue	Dataset access	Derived from...	Does the data have PII (personally identifiable information)?	Does the data have sensitive content/information?	Submission Date	Approval Status	Dataloader name	GitHub Issue URL
175	Indonesian Stance		ID Stance is a co	https://github.con	-	ind	javanese javanese kromo javanese ngoko dayak ahe dayak iban dayak pesaguan dayak taman buginese kelolau buginese wajo sundanese madurese banjar batak toba khek pontianak kapuas hulu melayu kembayan melayu ketapang melayu melawi melayu pontianak melayu putussibau melayu sambas melayu sintang padang tiociu pontianak	Stance Detection	Language	journalistic blog	Creative Commo	Crawling Expert-generatec				337	0	0		documents	337	Universitas Tanju	Stance Classifica	https://www.resea	2018		Free	Kompasiana	No	No		NusaCrowd carry-over	id_stance	
176	Indonesian WSD		Indonesian WSD	https://github.con	-	ind		Word Sense Disa	Language	multi domain	Unknown (unkno	Crawling				2,416	0	0		sentences	2416	Universitas Indor	Cross-Lingual an	https://aclantholo	2018		Free	CLWSD, WordNe				NusaCrowd carry-over	id_wsd	
177	Indonesian-Madurese Bible transla		The Madurese P	https://data.menc		mad ind	nline bible web pa	Machine Translat	Language	Religion	Creative Commo	Crawling	Automatic & Man						30013	sentence pairs	30013	State University c	Autogenerated In	https://papers.ssr	2023	Preprint	Free		No	No	2024-01-04 17:0	Approved	id_mad_bible	https://github.com
178	IndonesianNMT		Repository conta	https://github.con		ind jav min sun ban	N/A	Machine Translat	Language	General	Unknown (unkno	Machine-generat	Automatic						5000	sentence pairs	5000	Bandung Institute University of Indc Monash Universi Boston University	Replicable Bench	https://arxiv.org/a	2023	SEALP	Free	ChatGPT	No	No	2023-12-11 09:0	Approved	indonesiannmt	https://github.com
179	indonglish-dataset		Indonglish-datas	https://github.con	https://huggingfa	ind	N/A	Sentiment Analys	Language	Mixed	Unknown (unkno	Crawling Expert-generatec	Manual (full)	5		3638	399	1011		tweets	5048	FMIPA UGM Yog	Code-Mixed Sen	https://thesai.org/	2023	IJACSA	Free		No	No	2024-02-20 10:5	Approved	indonglish	https://github.com
180	IndoNLG Bible En-Id		Bible En-Id is a n	https://storage.gc	https://huggingfa	eng ind		Machine Translat	Language	religion	Creative Commo	Crawling Expert-translated				23,308	3,109	4,661		sentences	31078	Dyah Inastra	IndoNLG: Bench	https://aclantholo	2021		Free	Bible	No	No		NusaCrowd carry-over	bible_en_id	
181	IndoNLG Bible Jv-Id		Bible Jv-Id is a m the same split se	https://storage.gc	https://huggingfa	jav ind		Machine Translat	Language	religion	Creative Commo	Crawling Expert-translated				5,967	797	1,193		sentences	7957	Telkom University	IndoNLG: Bench	https://aclantholo	2021		Free	Bible				NusaCrowd carry-over	bible_jv_id	
182	IndoNLG Bible Su-Id		Bible Su-Id is a n	https://storage.gc	https://huggingfa	sun ind		Machine Translat	Language	religion	Creative Commo	Crawling Expert-translated				5,968	797	1,193		sentences	7958	Universitas Nege	IndoNLG: Bench	https://aclantholo	2021		Free	Bible	No	No		NusaCrowd carry-over	bible_su_id	
183	IndoNLI		IndoNLI is the firs	https://github.con	https://huggingfa	ind		Natural Languag	Language	news articles Wikipedia	Creative Commo	Expert-generatec				10,330	2,197	5,183		sentence pairs	17710	Carnegie Mellon	IndoNLI: A Natur	https://aclantholo	2021		Free	-	No	No		NusaCrowd carry-over	indonli	
184	IndoNLU NERGrit		NER Grit dataset	https://github.con	https://huggingfa	ind		Named Entiy Rec	Language	other	Creative Commo	Unknown				1,672	209	209		sentences	2090	Meta AI, UC Berk	IndoNLU: Bench	https://aclantholo	2020		Free	-	No	No		NusaCrowd carry-over	indonlu_nergrit	
185	IndoPuisi	Indonesian	Puisi is an Indon	https://github.con	https://huggingfa	ind		Language Model	Language	General	MIT (mit)	Crawling				7,223	0	0		documents	7223	Meta AI, UC Berk			2020		Free		No	No		NusaCrowd carry-over	indo_puisi	
186	IndoQA		IndoQA is an Indo	https://huggingfa	https://huggingfa	ind	N/A	Question Answer	Language	Mixed Multi-domain	Creative Commo	Expert-generatec	None			3310	1100	N/A		instances	4410	Jakarta AI Resea	N/A	https://huggingfa	2022	No publication	Free		No	No	2024-02-05 11:1	Approved	indoqa	https://github.com
187	IndoSMD		IndoSMD is a syı	https://github.con		ind	Indonesia	Dialogue System	Language	POI Navigation Weather Informatio Calendar Schedulin	Creative Commo	Crowdsourced	Manual (partial)						323	dialogues	323	Bandung Institute	IndoToD: A Multi-	https://arxiv.org/p	2023	SEALP	Free	SMD	Yes	No	2023-11-14 17:4	Approved	indosmd	https://github.com
188	IndoSRL	Indonesian	IndoSRL is a sen	https://github.con		ind eng		Semantic Role La	Language	General	Community Data	Expert-generatec				4,477	559	557		sentences	5593	Meta AI, UC Berk	Universal Propos	https://aclantholo	2022		Free	Universal Depen	No	No		NusaCrowd carry-over	indosrl	
189	IndoSum		The Indosum dat topics: entertainn	https://drive.goog	https://huggingfa	ind		Summarization	Language	news articles	Creative Commo	Crawling Expert-generatec				14,083	1,880	2,810		sentences	18773	Meta AI, UC Berk	Indosum: A new summarization.	https://arxiv.org/a	2021		Free	News aggregator	No	No		NusaCrowd carry-over	indosum	
190	IndoTacos		IndoTacos datase	https://www.kagg		ind		Legal Classificati	Language	Legal	Creative Commo	Crawling				12,291	0	0		documents	12291	Meta AI, UC Berk			2021		Free					NusaCrowd carry-over	indotacos	
191	IndoWiki		IndoWiki is a kno	https://github.con		ind	Indonesia	Knowledge Base	Language	Wikipedia	MIT (mit)	Crawling	None			4597264	26490	29636		examples	4653390	Pitik University of Indc	IndoKEPLER, Inc	https://ieeexplore	2022	International Wor	Free		Yes	No	2024-01-03 11:5	Approved	indowiki	https://github.com
192	IndQNER	Indonesian	IndQNER is a NE	https://github.con		ind		Named Entiy Rec	Language	Religion	Unknown (unkno	Expert-generatec				2,494	312	312		sentences	3118	Meta AI, UC Berk			2022		Free					NusaCrowd carry-over	indqner	
193	INDspeech_DIGIT_CDSR	Indonesian	INDspeech_DIGI	https://github.con		ind		Automatic Speec	Speech Language	General	Creative Commo	Expert-generatec				8,440	0	4,004		utterances [speal	12444	Meta AI, UC Berk	Indonesian Spee	https://www.isca-	2004		Free	-				NusaCrowd carry-over	indspeech_digit_	
194	INDspeech_NEWS_EthnicSR	Sundanese Javanese	INDspeech_NEW	https://github.con		sun jav		Automatic Speec	Speech Language	News articles	Creative Commo	Expert-generatec				2,000	0	300		utterances [speal	2300	Meta AI, UC Berk	Towards Languag	https://ieeexplore	2012		Free	-				NusaCrowd carry-over	indspeech_news_	
195	INDspeech_NEWS_LVCSR	Indonesian	INDspeech_NEW	https://github.con		ind		Automatic Speec	Speech Language	News articles	Creative Commo	Expert-generatec				39,600	0	4,400		utterances [speal	44000	Meta AI, UC Berk	Development of I	https://aclantholo	2008		Free					NusaCrowd carry-over	indspeech_news_	
196	INDspeech_NEWS_TTS	Indonesian	INDspeech_NEW	https://github.con		ind		Text-To-Speech S	Speech Language	News articles Travel expression	Creative Commo	Expert-generatec				1,972	0	40		utterances	2012	Meta AI, UC Berk	Development of I	https://www.sema	2008		Free					NusaCrowd carry-over	indspeech_news_	
197	INDspeech_NEWSTRA_EthnicSR	Indonesian Sundanese Javanese Balinese Bataks	INDspeech_NEW	https://github.con		ind sun jav ban btk		Automatic Speec	Speech Language	News articles Travel expression	Creative Commo	Expert-generatec				9,000	0	4,000		utterances [speal	13000	Samsul Rahmadi	Towards Languag	https://ieeexplore	2013		Free	-				NusaCrowd carry-over	indspeech_news	
198	INDspeech_TELDIALOG_LVCSR	Indonesian	INDspeech_TELI	https://github.con		ind		Automatic Speec	Speech Language	General Telephone applicati	Creative Commo	Expert-generatec				36,000	0	4,000		utterances [speal	40000	Indonesian-nlp	Development of I	https://aclantholo	2008		Free	-	No	No		NusaCrowd carry-over	indspeech_teldia	
199	INDspeech_TELDIALOG_SVCSR	Indonesian	INDspeech_TELI	https://github.con		ind		Automatic Speec	Speech Language	General Telephone applicati	Creative Commo	Expert-generatec				10,000	0	10,000		utterances [speal	20000	NAIST, ITB	Indonesian Spee	https://www.isca-	2004		Free	-	No	No		NusaCrowd carry-over	indspeech_teldia	
200	InSet Lexicon		Sentiment analys	https://github.con	-	ind		Sentiment Analys	Language	social media	Unknown (unkno	Crawling Expert-generatec				0	0	2,630		sentences	2630	R&D Division of I	InSet Lexicon: Ev	https://www.resea	2017		Free	-	No	No		NusaCrowd carry-over	inset_lexicon	
201	Inter-Agency Task Force for the Ma	iatfResolution52 iatfResolution20 iatfResolution46A iatfResolution35 iatfResolution01 iatfResolution56 iatfResolution33 iatfResolution09 iatfResolution40 iatfResolution39 iatfResolution21 iatfResolution10 iatfResolution44 iatfResolution43 iatfResolution08 iatfResolution22 iatfResolution51 iatfResolution23 iatfResolution58 iatfResolution36 iatfResolution32 iatfResolution11 iatfResolution60A iatfResolution27 iatfResolution53 iatfResolution19 iatfResolution30 iatfResolution48 iatfResolution47 iatfResolution59 iatfResolution38 iatfResolution14 iatfResolution03 iatfResolution31 iatfResolution15 iatfResolution12 iatfResolution60 iatfResolution50A iatfResolution26 iatfResolution57 iatfResolution41 iatfResolution07 iatfResolution25 iatfResolution30A iatfResolution02 iatfResolution55 iatfResolution55A iatfResolution29 iatfResolution24 iatfResolution42 iatfResolution05 iatfResolution46 iatfResolution54 iatfResolution34 iatfResolution37 iatfResolution45 iatfResolution04 iatfResolution28 iatfResolution17 iatfResolution49 iatfResolution18 iatfResolution06 iatfResolution50 iatfResolution13 iatfResolution16 fGuidelineOmnibu	This dataset cont Each resolution c	https://github.con		fil	Philippines	Language Model	Language	Health Legal	GNU General Pu	Crawling	None						90 94 154 146 53 154 113 69 160 83 158 78 62 177 96 168 84 140 58 96 114 137 175 92 60 157 70 156 151 88 656 231 55 92 165 136 84 113 60 103 173 80 81 59 87 57 170 105 124 82 84 79 153 128 55 102 72 108 151 106 108 80 65 204 87 753	sentences	8481	University of Oxf University of the I Cornell University	Creating text data	https://como-ph.g	2020	No publication	Free		No	No	2023-11-14 19:5	Approved	iatf	https://github.com
202	InterBEST-2009		InterBEST-2009 i	http://thailang.ne		tha	Thailand	Word lists	Language	Books News articles Encyclopedia	Creative Commo	Expert-generatec	Manual (full)			5000000	N/A	600000		tokens	5600000	NECTEC	BEST 2009 : Tha	https://ieeexplore	2009	International Sym	Free		No	No	2024-03-23 13:2	Approved with notes	interbest_2009	https://github.com
203	Intercontinental Dictionary Series (I		The Intercontiner	https://ids.clld.org		khb pcb prt cbn nut tha giq nod tyh smu san shn kkh mra blr syo tdd pnx cog zng thm bgk rbb sou vie	N/A	Word lists	Language	Dictionary	Creative Commo	Expert-generatec Expert-translated	None						N/A (at most 131	instances	0	Max Planck Instit	N/A		2023	No publication	Free	Various dictionari	No	No	2024-04-02 15:1	Approved	intercontinental_	https://github.com
204	JaDi-Ide	Standard Javanese Ngapak Javanese "East Javanese"	The dataset is co	https://github.con		jav		Dialect Identificat	Language	Social media	Unknown (unkno	Crawling Expert-generatec				16,000	0	0		tweets	16000	NAIST	Attention-based c	https://scholar.arc	2020		Free		No	No		NusaCrowd carry-over	jadi_ide	
205	JATI		JATI is a treeban for this study con official and the m sentences are pa 2015), a computa framework (Sag, knowledge is ext (Copestake et al.	https://github.con	-	ind		Constituency Par	Language	books	Unknown (unkno	Expert-generatec				1,253	0	0		sentences	1253	R&D Division of I	Building JATI: A Indonesian.	https://www.resea	2017		Free	KBBI	No	No		NusaCrowd carry-over	jati	
206	JV-ID ASR	Javanese	This data set con	http://openslr.org		jav		Automatic Speec	Speech Language	General	Creative Commo	Expert-generatec				185,076	0	0		utterances	185076	R&D Division of I	Crowd-Sourced S	http://dx.doi.org/1	2018		Free		No	No		NusaCrowd carry-over	jv_id_asr	
207	JV-ID TTS	Javanese	This data set con	http://openslr.org		ind jav		Text-To-Speech S	Speech Language	General	Creative Commo	Expert-generatec				5,800	0	0		sentences	5800	ATR Japan	A Step-by-Step P	http://dx.doi.org/1	2018		Free		No	No		NusaCrowd carry-over	jv_id_tts	
208	Kamus Alay		We provide a lex	https://github.con		ind		Morphological Inf	Language	Social media	Unknown (unkno	Expert-generatec Expert-translated				3,592	0	0		tokens	3592	R&D Division of I	Colloquial Indone	https://ieeexplore	2018		Free		No	No		NusaCrowd carry-over	kamus_alay	
209	Karonese dataset		Karonese datase	https://github.con		btx		Sentiment Analys	Language	Social media	Unknown (unkno	Crawling Expert-translated				1,001	0	0		sentences	1001	Google	Sentiment Analys	http://section.iaes	2022		Free		No	No		NusaCrowd carry-over	karonese_sentim	
210	KaWAT		We introduced Ka	https://github.con		ind		Word Analogy	Language	General	Apache license 2	Expert-generatec				34,000	0	0		sentence pairs	34000	Google	KaWAT: A Word A	https://arxiv.org/a	2019		Free		No	No		NusaCrowd carry-over	kawat	
211	KDE4		A parallel corpus	https://opus.nlpl.e		ind khm zlm tha vie	N/A	Machine Translat	Language	Software Technology	Unknown (unkno	Crowdsourced Expert-generatec Expert-translated	None						14.8k (ind) 100k (khm) 92.2k (zlm) 72.2k (tha) 43.6k (vie)	sentences	0	University of Hels Uppsala Universi	Parallel Data, Too	http://www.lrec-co	2012	LREC	Free		No	No	2024-01-21 06:0	Approved	kde4	https://github.com
212	KEPS		KEPS is a keyph	https://github.con	https://huggingfa	ind		Keyword Extracti	Language	banking	Creative Commo	Crawling Expert-generatec				1,000	0	247		documents	1247	Bina Nusantara U	Improving Joint L	https://www.sema	2019		Free	-	No	No		NusaCrowd carry-over	keps	
213	Kethu		Kethu is a constit	https://github.con	-	ind		Constituency Par	Language	news articles	Unknown (unkno	Machine-generat Expert-generatec				1,030	0	0		sentences	1030	Universitas Indor	Converting an Inc	https://colips.org/	2019		Free	ID-treebank-CSU				NusaCrowd carry-over	kethu	
214	Kheng.info Speech		The Kheng.info S	https://huggingfa	https://huggingfa	khm	Cambodia	Automatic Speec	Speech	Dictionary	Unknown (unkno	Crawling	None						3097	examples	3097	Independent	-		2023	No publication	Free		No	No	2024-01-21 13:0	Approved	kheng_info_spee	https://github.com
215	Khmer ALT POS		The data contain	https://www2.nict		khm	Cambodja	POS Tagging	Language	News	Creative Commo	Expert-generatec	None						20106	sentences	20106	National Institute	Towards Tokeniz	https://dl.acm.org	2021	ACM Transaction	Free		No	No	2023-11-16 18:2	Approved	khmer_alt_pos	https://github.com
216	khPOS		khPOS is a POS	https://github.con	https://huggingfa	khm	Cambodia	POS Tagging	Language	General	Creative Commo	Crawling Expert-generatec	Manual (full)			12000	1000	1000		sentences	14000	Waseda Universi Okayama Prefec National Institute	Comparison of Si	https://github.com	2017	Regional Confere	Free		Yes	No	2024-01-17 01:2	Approved	khpos	https://github.com
217	KoPI-CC (Korpus Perayapan Indon	Indonesia	KoPI-CC (Korpus	https://huggingfa	https://huggingfa	ind		Language Model	Language	Banking Books Commentary General Hotel reviews Journalistic blog News articles Religion Reviews Social media	Creative Commo	Crawling				106	0	0		GB	106	University of Can			2022		Free	Common Crawl				NusaCrowd carry-over	kopi_cc	
218	KoPI-CC_News	Indonesian	KoPI(Korpus Per	https://huggingfa	https://huggingfa	ind		Language Model	Language	News articles	Creative Commo	Crawling				4	0	0		GB	4	CMU			2022		Free	CC News	No	No		NusaCrowd carry-over	kopi_cc_news	
219	KoPI-NLLB	Indonesian Balinese nese,Minang,Jav	KopI(Korpus Pera each language se	https://huggingfa	https://huggingfa	ind sun jav min ban bjn ace		Language Model	Language	General Journalistic blog News articles Religion	Open Data Comr	Crawling				18	0	0		GB	18	Samsul Rahmadi			2022		Free		No	No		NusaCrowd carry-over	kopi_nllb	
220	Korpus Nusantara	Javanese Dayak Buginese Sundanese Madurese Banjar Batak Toba Khek Malay Minangkabau Tiociu	The dataset is a Since not all diale	https://github.con	https://github.com	ind jav xdy bug sun mad bjn bbc khek msa min tiociu		Machine Translat	Language	General	Unknown (unkno	Expert-translated				68,856	0	0		sentence pairs	68856	Samsul Rahmadi	Improving the rol	https://ijece.iaesc	2022		Free	-	No	No		NusaCrowd carry-over	korpus_nusantar	
221	KVIS Thai OCR		The KVIS Thai O styles of various	http://dx.doi.org/1		tha	Thailand	Optical Characte	Language Vision	General	Creative Commo	Crowdsourced	Manual (full)						1079	images	1079	Thai-Nichi Institu Kamnoetvidya Sc	Offline Handwritt	https://ieeexplore	2018	InCIT	Free		No	No	2023-12-24 01:1	Approved	kvis_th_ocr	https://github.com
222	Leipzig Corpora Collection		This is a collectio The 2017 Comm	https://wortschatz		ban bjn bew bcl mya ceb hil ind khm lao zsm min pam pag ksw tgl tha vie war jav mad	N/A	Language Model	Language	Mixed Multi-domain	Creative Commo	Crawling	None						N/A	examples	0	University of Leip	Building Large M	http://www.lrec-co	2012	LREC	Free		Yes	No	2023-12-15 15:1	Approved with notes	leipzig_corpora	https://github.com
223	LEX-INDO: AN INDONESIAN LEXI		This open-source	https://magichub.		ind	Asia	Lexical Normaliz	Language	General	Creative Commo	Expert-generatec	None						2000	tokens	2000	Beijing Magic Da	N/A		2021	No publication	Free		No	No	2024-02-11 06:1	Approved	lex_indo	https://github.com
224	Lexibank		Lexibank is a pub	https://github.con		fau jeh nia tmu wrs	N/A	Word lists	Language	Dictionary	Creative Commo	Expert-generatec Machine-generat Expert-translated	Automatic						709,638 lexeme 3,857,425 tokens	instances	0	Max Planck Instit	Lexibank, a publi	https://www.natur	2022	Nature	Free	Compilation from	No	No	2024-04-02 14:5	Approved	lexibank	https://github.com
225	LEXiTRON	version 2.0	Corpus-based di	https://opend-por		tha eng	Thailand	Word-level Trans Machine Translat	Language	Multi-domain Bilingual Dictionary	Custom NECTEC	Machine-generat Expert-translated	Manual (full)						83000 53000	instances	136000	NECTEC	Qualitative and C	https://www.sema	2002	COCOSDA	Free		No	No	2024-03-10 16:1	Approved with notes	lexitron	https://github.com
226	LibriVox-Indonesia	ace bal bug id min jav sun (for Acehnese Balinese Bugisnese Indonesian Minangkabau nese and Sundan	The LibriVox Indo	https://huggingfa	https://huggingfa	ind sun jav min bug ban ace		Automatic Speec	Speech Language	Books Transcribed audio Declaration of Hun	Creative Commo	Crawling				7,815	0	0		utterances	7815	Samsul Rahmadi	LibriVox Indonesi	https://huggingfa	2022		Free	https://librivox.org	No	No		NusaCrowd carry-over	librivox	
227	LimeSoda		Thai fake news d	https://github.con		tha	N/A	Fact Checking	Language	Health	Creative Commo	Expert-generatec	Manual (full)			2698	300	2765		documents	5763	Backyard.co	LimeSoda: Datas	https://ieeexplore	2021	ISAI-NLP	Free		No	No	2023-11-17 15:0	Approved	limesoda	https://github.com
228	Lio and the Central Flores languag	ASR-end ASR-nxe ASR-ssq ASR-ljl MT-eng-end MT-eng-nxe MT-eng-ssq MT-eng-ljl	This dataset is a	https://archive.mp		end nxe ssq ljl eng	Ende	Automatic Speec Machine Translat	Language Vision Speech	Conversational Education Mixed Folklore	Unknown (unkno	Expert-generatec	Manual (partial)						50 38 69 460 50 38 69 460	hours	0	Leiden University	Lio and the Centr	https://studentthe	2018	Preprint	Free		Yes	Yes	2024-01-07 15:0	Approved with notes	lio_and_central_f	https://github.com
229	Liputan6 Summarization		The Liputan6 dat	https://storage.gc	https://huggingfa	ind		Summarization	Language	news articles	Creative Commo	Crawling Expert-generatec				193,883	10,972	10,972		sentences	215827	Tallinn University	Liputan6: A Large	https://aclantholo	2021		Free	Liputan6 online I				NusaCrowd carry-over	liputan6	
230	Local ID Abusive	Javanese Sundanese	Local ID Abusive	https://github.con		sun jav		Hate Speech Det	Language	Social media	Unknown (unkno	Crawling Expert-generatec				5,656	0	0		sentences	5656	University of Pad	Abusive languag	http://www.wcse.	2021		Free		No	No		NusaCrowd carry-over	local_id_abusive	
231	LR-Sum		LR-Sum is a new	https://huggingfa	https://huggingfa	ind vie lao tha khm mya	N/A	Abstractive Sumr	Language	News articles	Creative Commo	Crawling	None			968 11676 11964 2278 3888 7921	500 1459 1495 500 486 990	500 1460 1496 500 486 990		article-summary p	49557	Brandeis Univers	LR-Sum: Summa	https://aclantholo	2023	ACL	Free	Multilingual Oper	No	No	2024-01-17 01:0	Approved	lr_sum	https://github.com
232	LSVSC		A large-scale Vie	https://drive.goog		vie	Vietnam	Automatic Speec	Language Speech	Books Health Culture & heritage News Sports Tourism Movies/TV Traffic Audio books	Creative Commo	Crawling Crowdsourced	Manual (full)			45458	5682	5683		utterances	56823	Posts and Teleco NAVER corporati	Automatic Speec	https://www.mdpi	2023	Electronics (MDF	Free		No	No	2024-03-25 22:3	Approved	lsvsc	https://github.com

No	Dataset name	Subset(s)	Dataset description	Dataset URL	HuggingFace URL	Dataset language(s)	Dataset collection region	Dataset task(s)	Dataset modality	Dataset domain(s)	Dataset license	Dataset annotation collection style	Dataset annotation validation style	Number of annotators per sample	Inter-annotator agreement	Train split data size	Validation split data size	Test split data size	No split data size	Data size unit	Total data size	Dataset provider	Dataset paper title	Dataset paper URL	Dataset or dataset paper publish year	Publication venue	Dataset access	Derived from...	Does the data have PII (personally identifiable information)?	Does the data have sensitive content/information?	Submission Date	Approval Status	Dataloader name	GitHub Issue URL
233	LTI LangID Corpus		The LTI LangID c	https://www.cs.c		ifa ace btm mqj mbs nbq gor zyp tbl kjp kmk kqe ptu blw ceb prf yva zlm bps tdt mya dgc lus wrs abx rgu aaz agn ccp cmn jav obo due msk xsb syb ind lbk min smk att nod tdj atb atd pag hvn ksc lao kkl lti dao cgc tbk gdg amk mbd clu msb sbl cek khm yue sgb beu eip ifu mnw suc bgs pam ebk eng nfa cbk ify csy heg shn mta mbt lex tha mmn sun vie llg xnn txq bcl kje kne san hlt kyu bkd duo tet ury yka bjn tiy ivv agt ban blz mbi ilo mkn isd cth bpr por mbb tgl msm ivb tam plw alp row	N/A	Language Identif Language Model	Language	Religion Wikipedia Politics	Other (other)	Crawling Expert-generatec Expert-translated	None			750	25	25		MB	800	Carnegie Mellon	Non-linear Mappi	https://aclantholo	2014	EMNLP	Free		No	No	2024-01-21 05:3	Approved	lti_langid_corpus	https://github.com
234	M3Exam	Javanese Thai Vietnamese	M3Exam is a nov	https://github.con		jav tha vie	esia, Thailand, Vi	Question Answer Visual Question /	Language Vision	School Exam	Creative Commo	Crowdsourced	Manual (full)			0	0	382 2595 1933		instances	4910	The Chinese Uni Alibaba Group SUTD	M3Exam: A Multi	https://arxiv.org/a	2023	NeurIPS	Free		No	No	2023-11-24 16:5	Approved	m3exam	https://github.con
235	M3IT	imagenet msrvtt-qa msrvtt okvqa vist vqa-v2 winoground	The Multimodal,	https://huggingfa	https://huggingfa	eng ceb ind jav khm lao mya tha tgl vie	N/A	Commonsense R Visual Storytelling Question Answer Image-to-Text Ge	Language Vision	Conversational Education Health Social media News environment Birds Attraction Taxi Train airlines Academic Sports Animals Science Food Ads	Other (other)	Machine-translat	Automatic			500 500 500 500 500 500 0	500 497 500 500 500 500 0	0 500 500 0 500 0 800		images	8297	The University of Peking University	M3IT: A Large-Sc	https://arxiv.org/a	2023	Preprint	Free	ImageNet, VQA v	No	No	2023-12-25 14:4	Approved	m3it	https://github.con
236	M3LS		The multilingual	https://github.con		ind	N/A	Multimodal Sumr	Language Vision	News articles News	MIT (mit)	Crawling Expert-generatec	Manual (partial)			50498	2805	2805		document-summ	56108	Indian Institute of	Large Scale Mult	https://aclantholo	2023	EACL	Free	BBC	No	No	2023-12-26 04:2	Approved	m3ls	https://github.con
237	MABL	ind jav sun	The MABL (Meta	https://github.con	https://huggingfa	ind jav sun	Indonesia	Question Answer	Language	Metaphor	MIT (mit)	Crowdsourced	Manual (full)						1140 600 600	sentence pairs	2340	MBZUAI HKUST Carnegie Mellon Bloomberg Masakhane	Multi-lingual and	https://aclantholo	2023	ACL	Free		No	No	2023-11-16 16:1	Approved	mabl	https://github.con
238	MadureseSet		MadureseSet is a	https://data.menc		mad	Madura	POS Tagging Word Sense Disa Machine Translat Word lists Dialect Identificat	Language	Dictionary	Creative Commo	Machine-generat	Automatic & Man						17809 53722	relations	71531	multiple institutio	MadureseSet: Ma	https://www.scien	2023	Elsevier Sciencel	Free	The Complete Di	No	No	2023-11-14 13:1	Approved	madurese_set	https://github.con
239	Malaysia-AI government	gov.my.jsonl govdocs.jsonl uftiwp.gov.my.jso ns.mohe.gov.my.j	This is a dataset	https://huggingfa		zlm	Malaysia	Language Model	Language	General Multi-domain Religion Web Politics National Policy	Apache license 2	Crawling	None	0	0				31031	documents	31031	Independent Malaysia-AI	N/A		2023	No publication	Free		No	No	2024-01-20 22:3	Approved	malaysia_ai_gov	https://github.con
240	Malaysia-AI Hansard Scrape		The Malaysia AI	https://huggingfa	https://huggingfa	zlm	Malaysia	Language Model	Language	Legal Web Politics	Apache license 2	Crawling	None						142776	documents	142776	Independent Malaysia-AI	N?A		2023	No publication	Free		No	No	2024-01-30 00:0	Approved	malaysia_ai_han	https://github.con
241	malaysia-tweets-with-sentiment-lab		This tweet data w	https://github.con		zlm eng	Malaysia	Sentiment Analys	Language	Health	Unknown (unkno	Crawling Machine-generat	None						6999	tweets	6999	Faculty of Compu	Social Versus Ph	https://link.spring	2022	International Con	Free		No	No	2024-02-11 10:4	Approved	malaysia_tweets	https://github.con
242	MALINDO_Morph	Kamus Dewan Besar Bahasa Inc zig Corpora Collec Frogstory-David layu-Standard-Lis Melayu-Sabah Melayu-Sarawak Melayu-Brunei ndo-Jakarta-Lisar	MALINDO Morph	https://github.con		zlm ind	N/A	Morphological Inf	Language	General	Creative Commo	Machine-generat	Manual (full)						246525	tokens	246525	Nanyang Techno Tokyo University	MALINDO Morph	http://lrec-conf.or	2018	LREC	Free		No	No	2023-12-16 14:1	Approved	malindo_morph	https://github.con
243	MALINDO_parallel		Teks ini adalah s	https://github.con		zlm ind	N/A	Machine Translat	Language	Conversational General	Creative Commo	Expert-generatec	None						86	sentence pairs	86	Tokyo University	N/A	https://github.con	2020	No publication	Free		Yes	No	2023-12-16 13:3	Approved	malindo_parallel	https://github.con
244	MaRVL	Indonesian	Multicultural Rea	https://marvl-chal		ind		Image-to-Text Ge	Language	General	Creative Commo	Crawling Expert-generatec				1,128	0	0		image, image, co	1128	IR-NLP Lab Fasil	Visually Grounde	https://arxiv.org/p	2021		Free		No	No		NusaCrowd carry-over	marvl	
245	MASSIVE	id-ID jv-ID km-KH ms-MY my-MM th-TH tl-PH vi-VN	MASSIVE datase contains 1M reali	https://github.con	https://huggingfa	ind jav khm zlm mya tha tgl vie cmn tam	N/A	Slot Filling Intent Classificati	Language	Multi-domain	Creative Commo	Expert-generatec	Manual (full)			11514	2033	2974		utterances	16521	Amazon	MASSIVE: A 1M-	https://aclantholo	2023	ACL	Free	SLURP	No	No	2023-11-17 05:0	Approved	massive	https://github.con
246	MassiveSumm	fil ind khm lao mya tha vie	A (massive) multi	https://github.con		fil ind khm lao mya tha vie	N/A	Language Model Summarization	Language	Mixed	Unknown (unkno	Crawling Crowdsourced	Manual (partial)						N/A	article-summary	0	IT University of C	MassiveSumm: a	https://aclantholo	2021	EMNLP	Free		No	No	2024-03-16 00:5	Approved	massivesum	https://github.con
247	MaXM	MaXM v1 -th	MaXM, a test-onl	https://github.con		tha	SEA	Visual Question /	Language Vision	Mixed	Other (other)	Expert-generatec Machine-translat	Manual (partial)	7		0	0	302		instances	604	Google	MaXM: Towards		https://aclantholo	2023	EMNLP Finding	Free	XM3600	No	No	2024-02-01 23:2	Approved	maxm	https://github.con
248	mc4-indo		A thoroughly clea	https://huggingfa	https://huggingfa	ind	N/A	Language Model	Language	General Mixed Web	Open Data Comr	Crawling	Automatic						N/A	paragraphs	0	Independent Google	m{T}5: A Massive	https://aclantholo	2021	ACL	Free		No	No	2023-11-11 21:38	Approved	mc4_indo	https://github.con
249	MDIA	ind_dialogue ind_eng tgl_dialogue tgl_eng vie_dialogue vie_eng	This is a multiling	https://github.con		ind tgl vie	N/A	Dialogue System	Language	Social media	Creative Commo	Crawling Machine-translat	Manual (partial)			10000 20000 10000 20000 2119 4238	1000 2000 1000 2000 212 424	1000 2000 1000 2000 1000 2000		examples	81993	Saarland Univers Nanjing Universit	MDIA: A Benchm	https://arxiv.org/p	2022	arXiv	Free		No	No	2024-03-14 13:2	Approved	mdia	https://github.con
250	MedEV		A high-quality Vie medical domain,	https://huggingfa	https://huggingfa	vie	Vietnam	Machine Translat	Language	Health	Unknown (unkno	Expert-translated Machine-translat	Automatic			340897	8982	9006		sentence pairs	358885	VinAI Research University of Tec VinUniversity	Improving Vietna	https://arxiv.org/p	2024	LREC	Free		No	No	2024-04-01 12:3	Approved	medev	https://github.con
251	MEDISCO	Train Test	MEDISCO is a M	https://huggingfa	https://huggingfa	ind	Indonesia	Automatic Speec	Speech	Health	GNU General Pu	Crawling Expert-generatec	Manual (full)			8.52	N/A	1.5		hours	10.02	Universitas Indor	Building MEDISC	https://ieeexplore	2018	IALP	Free		No	No	2024-03-31 23:1	Approved	medisco	https://github.con
252	Melayu_Brunei		This article gives	https://github.con		zlm kxd	Weston	Language Model	Language	Education General Folklore	Creative Commo	Expert-generatec	None						Contact me for d	tokens	0	Independent Tokyo University	Two Brunei Mala	https://tufs.repo.n	2021	Asian and Africar	Free		No	No	2023-12-20 17:3	Approved	melayu_brunei	https://github.con
253	Melayu_Sabah		Korpus Variasi Ba	https://github.con		msi	Sandakan Tenom Weston Sandakan Semporna Tambunan	Language Model	Language	Education General Folklore	Creative Commo	Expert-generatec	None						Contact me for d	tokens	0	Tokyo University Universiti Kebang	N/A		2018	No publication	Free		No	No	2023-12-19 17:5	Approved	melayu_sabah	https://github.con
254	Melayu_Sarawak		Korpus Variasi Ba	https://github.con		zlm	Sarawak Asajaya	Language Model	Language	Education General Folklore	Creative Commo	Expert-generatec	None						Contact me for d	tokens	0	Tokyo University	N/A		2019	No publication	Free		No	No	2023-12-20 01:0	Approved	melayu_sarawak	https://github.con
255	Melayu_Standard_Lisan		Korpus Variasi Ba	https://github.con		zlm	Johor Kuala Lumpur Negeri Sembilan Sabah	Language Model	Language	Education Culture & heritage Talk Folklore	Creative Commo	Expert-generatec	None						Contact me for d	tokens	0	Tokyo University	N/A		2018	No publication	Free		No	No	2023-12-19 17:4	Approved	melayu_standard	https://github.con
256	MEmoLon	ind zlm tgl vie sun tha mya ceb jav	MEmoLon is an e	https://zenodo.or		ind zlm tgl vie sun tha mya ceb jav	N/A	Emotion Classific	Language	Multi-domain	MIT (mit)	Machine-generat	Automatic						Indonesian - 2,0(Malay - 1,213,39 Vietnamese - 2,0 Javanese - 518,6 Sundanese - 327 Thai - 2,006,540 Cebuano - 2,006	tokens	0	Jena University L	Learning and Eva	https://aclantholo	2020	ACL	Free		No	No	2023-11-08 12:0	Approved	memolon	https://github.con
257	MEN-Dataset		The Malaysian E	https://github.con		eng	Malaysia	Named Entity Re Relation Extractio	Language	News articles	MIT (mit)	Crawling Expert-generatec	Manual (full)	2	0.818				6061 3268	instances	9329	Monash Universi Valiantlytix Sdn B	Malaysian Englis	https://arxiv.org/a	2024	LREC-COLING	Free		No	No	2024-03-25 18:0	Approved	men	https://github.con
258	MinangNLP MT		MinangNLP MT i	https://github.con	-	ind min		Machine Translat	Language	wikipedia	MIT (mit)	Crawling Expert-generatec				11,571	1,600	3,200		sentences	16371	IBM Research	Towards Comput	https://aclantholo	2020		Free	-	No	No		NusaCrowd carry-over	minangnlp_mt	
259	MinangNLP Sentiment Analysis		We release the M	https://github.con		ind min	N/A	Sentiment Analys	Language	Mixed	MIT (mit)	Crowdsourced	Manual (full)						5000	documents	5000	The University O	Towards Comput	https://aclantholo	2020	PACLIC	Free		No	No	2024-03-15 20:2	Approved	minang_senti	https://github.con
260	MIRACL	MIRACL Indonesia MIRACL Thai	MIRACL is a mul where all annotat	https://project-mi	https://huggingfa	ind tha	N/A	Text Retrieval	Language	Wikipedia	Apache license 2	Expert-generatec	Automatic & Man			4071 2972	960 733	1342 1642		instances	11720	University of Wat	MIRACL: A Multil	https://direct.mit.e	2023	TACL	Free		No	No	2023-11-08 06:1	Approved	miracl	https://github.con
261	MKQA	KQA Khmer (km-K KQA Malay (ms-M MKQA Thai (th-TH A Vietnamnese (v	MKQA, an open-	https://github.con		khm zsm tha vie	N/A	Question Answer	Language	Wikidata	Creative Commo	Expert-generatec	Manual (full)			0 0 0 0	0 0 0 0	10000 10000 10000 10000		instances	40000	Apple	MKQA: A Linguis	https://direct.mit.e	2021	TACL	Free	Google Natural C	No	No	2023-11-08 06:0	Approved	mkqa	https://github.con
262	MLQA	lqa-translate-test. lqa-translate-train. mlqa.vi.ar mlqa.vi.de mlqa.vi.vi mlqa.vi.zh mlqa.vi.en mlqa.vi.es mlqa.vi.hi mlqa.ar.vi mlqa.de.vi mlqa.zh.vi mlqa.en.vi mlqa.es.vi mlqa.hi.vi	MLQA (MultiLing	https://github.con	https://huggingfa	vie	N/A	Question Answer	Language	Wikipedia	Creative Commo	Crawling Crowdsourced Expert-generatec	Manual (partial)	3	0.84	132600	0	0		instances	132600	Meta	MLQA: Evaluatin	https://aclantholo	2020	ACL	Free	Wikipedia	No	No	2023-10-25 16:2	Approved	mlqa	https://github.con
263	MLQE-PE	Khmer	We present MLQ	https://github.con https://github.con		khm	N/A	Machine Translat	Language	Mixed	Creative Commo	Crowdsourced	Manual (partial)						990	examples	990	Johns Hopkins U	MLQE-PE: A Mul	https://aclantholo	2022	LREC	Free		No	No	2024-03-16 01:3	Approved	mlqe_pe	https://github.con
264	MM-Sum		MM-Sum is the fi	https://drive.goog		ind vie mya tha	N/A	Multimodal Sumr	Language Vision	News	Unknown (unkno	Crawling	None			Contact me for d	Contact me for d	Contact me for d		article-summary	0	WeChat AI Beijing Jiaotong U	Summary-Orient	https://aclantholo	2023	ACL	Free	XL-Sum	No	No	2024-02-17 10:3	Approved with notes	mm_sum	https://github.con
265	MongabayConservationDataset	Multi-label ntiment-classificat	Conservation dat	https://huggingfa	https://huggingfa	ind	a Island, Indones	Sentiment Analys Topic Modeling	Language	News articles environment	The Unlicense (u	Machine-generat	Automatic & Man	1		3919	492	485		paragraphs	4896	Datasaur.ai	Utilizing Weak Su	https://arxiv.org/p	2023	SEALP	Free	https://www.mon	No	No	2023-11-14 19:3	Approved	mongabay	https://github.con
266	Mozilla Pontoon Translations	Burmese - English Cebuano - English Gorontalo - English Hiligaynon - English Ilocano - English Indonesian - English Javanese - English Khmer - English Lao - English Malay - English Nias - English Tagalog - English Thai - English etnamnese - Engli	This is a dataset https://pontoon.m	https://huggingfa	https://huggingfa	mya ceb gor hil ilo ind jav khm lao zlm nia tgl tha vie	multiple	Machine Translat	Language	General Software	BSD 3-clause "N	Crowdsourced	None						14800 686 278 361 1860 45100 708 15900 16100 31100 186 15600 38400 44500	sentence pairs	225579	Mozilla	n/a		2011	No publication	Free		No	No	2023-12-22 17:4	Approved	mozilla_pontoon	https://github.con
267	MSVD-Indonesian		MSVD-Indonesia	https://github.con		ind	Indonesia	Text Retrieval Image-to-Text Ge	Language Vision Video	Multi-domain	MIT (mit)	Machine-generat	Manual (partial)			1200	100	670		instances	1970	Independent	MSVD-Indonesia	https://arxiv.org/a	2023	Preprint	Free	Microsoft Video D	No	No	2023-11-15 22:0	Approved	id_msvd	https://github.con
268	MTOB	Grammar Book Bilingual Wordlist allel Sentence Co	The Machine Tra	https://github.con		kgv eng	llage, Indonesian	Machine Translat Language Model Word lists	Language	Conversational General Mixed Multi-domain Culture & heritage Stimuli	Creative Commo	Expert-generatec	Manual (full)			400	N/A	100		sentence pairs	500	Google Lund University Stanford Universi University of Oslo Oxford University	A Benchmark for	https://arxiv.org/a	2024	Preprint	Free upon reque	A grammar of Ka	Yes	No	2024-02-29 09:2	Approved	mtob	https://github.con
269	MTOP Intent Classification	Domain Intent	An almost-paralle	https://fb.me/mto	https://huggingfa https://huggingfa	tha	N/A	Constituency Par Semantic Role La Intent Classificati	Language	Conversational	Unknown (unkno	Crowdsourced Expert-generatec Machine-generat	None			10759 13400	1671 1820	2765 3550		sentences	37540	Meta	MTOP: A Compre	https://arxiv.org/a	2021	ACL	Free		No	No	2023-11-16 02:0	Approved	mtop_intent_clas	https://github.con
270	MultiJail		The MultiJail data	https://github.con	https://huggingfa	jav vie tha	N/A	Language Model	Language	Legal Ethics	MIT (mit)	Expert-generatec	Manual (full)						315	instances	315	The Chinese Uni Alibaba Group Nanyang Techno	Multilingual Jailb	https://arxiv.org/a	2023	Preprint	Free		No	Yes	2023-11-16 16:3	Approved	xl_jailbreak	https://github.con
271	Multilabel Multiclass Sentiment and		This dataset cont	https://github.con		ind	N/A	Sentiment Analys Emotion Classific	Language	Reviews Technology	Creative Commo	Crawling	Manual (full)	2	0.6114				21697	comments	21697	Bina Nusantara U	Multilabel multicla	https://www.scien	2023	Data in Brief	Free		No	No	2024-01-02 07:2	Approved	id_sent_emo_mc	https://github.con

No	Dataset name	Subset(s)	Dataset description	Dataset URL	HuggingFace URL	Dataset language(s)	Dataset collection region	Dataset task(s)	Dataset modality	Dataset domain(s)	Dataset license	Dataset annotation collection style	Dataset annotation validation style	Number of annotators per sample	Inter-annotator agreement	Train split data size	Validation split data size	Test split data size	No split data size	Data size unit	Total data size	Dataset provider	Dataset paper title	Dataset paper URL	Dataset or dataset paper publish year	Publication venue	Dataset access	Derived from...	Does the data have PII (personally identifiable information)?	Does the data have sensitive content/information?	Submission Date	Approval Status	Dataloader name	GitHub Issue URL
272	Multilanguage Open Text (MOT)		The Multilanguag From the paper: "Each file contair • filename: the na URL • url: the URL fro • url origin: the si was retrieved • content type: th • site language: t • time published: document was pu • time modified: t document was la • time retrieved: t document was re • title: the title of t • authors: the aut • paragraphs: the • n paragraphs: t the document • n chars: the nur • cld3 detected la extracted text of t – language: the l CLD3 – probability: the from CLD3) – is reliable: if pr (passed directly f – proportion: the identified as the l from CLD3) • predicted langu predict that the d that take into acc by CLD3 • keywords: the t of the document • section: the sub These additional languages: • sentences: the t • n sentences: th document • tokens: the text segmented into t • n tokens: the nu • parallel english the English docu (this currently onl	https://github.con		lao tha vie ind khm mya	VOA Websites	Language Model	Language	Commentary News articles Websites Snippets	MIT (mit)	Crawling Machine-generat	Automatic						Contact me (or ic	sentences	0	Brandeis Univers	Multilingual Oper	https://aclantholo	2022	LREC	Free		No	No	2024-01-13 23:1	Approved	mot	https://github.con
273	MultiLexNorm	ian-English (code	MultiLexNorm is	https://bitbucket.c		ind eng		Lexical Normaliz	Language	Social media	Creative Commo	Crawling Expert-generated				13,950	4,810	4,367		tokens	23127	Opensubtitles	MultiLexNorm: A	https://aclantholo	2021		Free		No	No		NusaCrowd carry-over	multilexnorm	
274	Multilingual Open Relations		This dataset cont	https://www.kagg		ind zlm tha vie tgl khm lao	N/A	Relation Extractio	Language	Wikipedia	Creative Commo	Machine-generat	None						N/A	relations	0	Carnegie Mellon	Multilingual Oper	https://aclantholo	2015	NAACL	Free		No	No	2024-03-28 20:5	Approved	multilingual_oper	https://github.con
275	Multilingual Open Relations	Indonesian	This dataset prov	https://www.kagg		ind		Relation Extractio	Language	Wikipedia	Unknown (unkno	Machine-translat				1,876	0	0		relations	1876	Jim Geovedi / Ar	Multilingual Oper	https://aclantholo	2015		Free		No	No		NusaCrowd carry-over	open_relation_id	
276	Multilingual Spoken Words Corpus	cnh_wav ind_wav vie_wav cnh_opus ind_opus vie_opus	Multilingual Spok	https://huggingfa	https://huggingfa	cnh ind vie	Unknown	Automatic Speec	Language Speech	Conversational	Creative Commo	Machine-generat	None			5750 28800 215 5750 28800 215	935 4420 73 935 4420 73	933 4360 62 933 4360 62		utterances	45548	SIL International	Multilingual Spok	https://datasets-b	2021	NeurIPS	Free		No	No	2023-11-27 10:4	Approved	mswc	https://github.con
277	Multilingual-ALPACA		For multilingual g The six language	https://github.con		vie	NA	Chatbot	Language	Mixed	Unknown (unkno	Machine-translat	Automatic						52000	examples	52000	Nanjing Universit	EXTRAPOLATIN	https://arxiv.org/p	2023	Preprint	Free	ALPACA	No	No	2024-03-14 17:3	Approved	multilingual_alpa	https://github.con
278	multilingual-NLI-26lang-2mil7		This dataset cont	https://huggingfa	https://huggingfa	ind vie eng	N/A	Natural Languag	Language	Mixed	Unknown (unkno	Crawling Crowdsourced	Automatic						varies on langua	examples	0	Vrije Universiteit	Less Annotating,	https://www.camb	2023	Political Analysis	Free		No	No	2024-03-16 01:4	Approved with notes	multilingual_nli_2	https://github.con
279	multispider	multispider_vi	MULTISPIDER, t seven languages Chinese, and Vie identify the lexica (caused by speci and their intensity	https://github.con	https://huggingfa	vie	N/A	Text-to-SQL	Language	Education Wikipedia	Creative Commo	Machine-generat	Automatic & Man	15		8659	1034	N/A		examples	9693	Harbin Institute o	MultiSpider: Towa	https://arxiv.org/p	2023	AAAI Conference	Free	Spider Text-to-SQ	No	No	2023-12-29 08:0	Approved	multispider	https://github.con
280	MUSE (Multilingual Unsupervised a		Contains 110 larg	https://github.con		fil ind zlm tha vie	N/A	Machine Translat Word lists	Language	General Dictionary	Creative Commo	Machine-generat	None			5000	0	1500		instances	6500	Meta	Word Translation	https://openrevie	2018	ICLR	Free		No	No	2024-01-02 12:1	Approved	muse	https://github.con
281	Myanmar (Burmese) Name Roman		This dataset cons	http://www.nlpres		mya	Myanmar	Machine Translat	Language	Names	Creative Commo	Expert-generated	Manual (full)						2,335	examples	0	University of Con	Statistical Romar	https://att-astrec.	2017	The Association f	Free		No	No	2023-11-15 05:3	Approved	burmese_romani	https://github.con
282	Myanmar Part-of-Speech (myPOS)		This version of th The format used All new raw sente	https://github.con		mya	Myanmar	POS Tagging	Language	Mixed Multi-domain Wikipedia	Creative Commo	Crawling	Automatic & Man			42196	0	1000		sentences	43196	King Mongkut's I National Electron	Myanmar POS R	https://ieeexplore	2020	Institute of Electr	Free	Asian Language	No	No	2023-11-15 08:5	Approved	mypos	https://github.con
283	Myanmar Sign Language Corpus fo		The MSL4Emerg	https://github.con		ysm mya	Myanmar	Machine Translat Sign Language R	Language Vision	Emergencies	Creative Commo	Expert-generated	Automatic & Man						558	instances	558	University of Tec Yangon Technolo Waseda Universi Okayama Prefec	Statistical Machir	https://core.ac.uk	2018	ICCA	Free		No	No	2023-12-24 07:4	Approved	msl4emergency	https://github.con
284	Myanmar-Rakhine Parallel Data		The data contain	https://github.con		mya rki	N/A	Machine Translat	Language	travel	GNU General Pu	Expert-generated	Manual (full)			14076	2485	1812		sentences	18373	Waseda Universi University of Con	Neural Machine T	https://aclantholo	2019	ACL Workshop	Free		No	No	2023-12-25 20:5	Approved	myanmar_rakhin	https://github.con
285	myParaphrase		The myParaphra	https://github.con		mya	Myanmar Thailand Cambodia	Paraphrasing	Language	General Social media	Creative Commo	Crawling Expert-generated	Manual (full)			40461	0	open-test: 1000 closed-test: 1000		sentence pairs	40461	National Electron Institute of Digital Cambodia Acade University of Tec	Deep Siamese N	https://jiist.aiat.or	2022	JIIST	Free		No	No	2023-12-26 20:1	Approved	my_paraphrase	https://github.con
286	mySentence	sentences ntences+paragrap	mySentence is a	https://github.con		mya	Myanmar	POS Tagging	Language	Multi-domain	Creative Commo	Crawling Expert-generated	Automatic & Man			sentences: 4000(sentences+parag	sentences: 2414 sentences+parag	sentences: 4712 sentences+parag		examples	0	National Electron	mySentence: Se	https://ph05.tci-th	2023	JIIST	Free		No	No	2023-12-26 13:3	Approved	mysentence	https://github.con
287	MyWSL2023		This dataset cont	https://data.menc		xml	Malaysia	Sign Language R	Vision	Dictionary	Creative Commo	Expert-generated	None			3000	0	500		images	3500	Universiti Kebang UCSI University	MyWSL: Malaysia	https://www.scien	2023	Data in Brief	Free		No	No	2024-01-01 07:0	Approved	mywsl2023	https://github.con
288	myXNLI		The myXNLI corp	https://github.con	https://huggingfa	mya	Myanmar	Natural Languag	Language	Multi-domain	Creative Commo	Expert-generated Machine-generat	Manual (partial)			392702	1490	5010		sentence pairs	6500	Independent	-		2023	No publication	Free	XNLI and MultiNL	No	No	2023-12-26 15:4	Approved	myxnli	https://github.con
289	National Speech Corpus	read_balanced read_pertinent conversational_f2 versational_teleph	The National Spe	https://www.imda		eng	Singapore	Automatic Speec Text-To-Speech S	Language Speech	Multi-domain	Other (other)	Expert-generated	Manual (full)						1258291.2	MB	1258291.2	IMDA Singapore	National Speech	https://docs.goog	2018	Preprint	Free upon reque		No	No	2024-03-14 14:3	Approved	national_speech_	https://github.con
290	Native Ads Dataset		The native ads d	https://figshare.c		ind	Indonesia	Text Classificatio	Language	News	Creative Commo	Crawling	Manual (full)						9354	documents	9354	Institut Teknologi	A Comprehensive	https://doi.org/10.	2023	2023 IEEE 8th In	Free		No	No	2024-04-03 17:3	Approved	ind_native_ads_(	https://github.con
291	NECID	ve Content Identif	NECID provides	https://github.con		ind	Indonesia	Text Classificatio	Language	Tweet	Creative Commo	Crawling	Automatic & Man			10959	2192	5114		tweets	18265	Universitas Islam	Fine-Tuning BER	https://ieeexplore	2023	IEEE	Free		No	Yes	2024-05-13 00:0	Approved	necid	https://github.con
292	NER UGM (IndoLEM split)		NER UGM is a n	https://github.con	-	ind		Named Entiy Rec	Language	news articles	Creative Commo	Crawling Expert-generated				1,687	187	469		sentences	2343	School of Electric	Name Entity Rec	http://etd.reposito	2014		Free	-	No	No		NusaCrowd carry-over	indolem_ner_ugr	
293	NER UI (IndoLEM split)		NER UI is a nam	https://github.con	-	ind		Named Entiy Rec	Language	general	Creative Commo	Expert-generated				1,530	170	425		sentences	2125	Universitas Islam	Automatic open c	https://ieeexplore	2017		Free	-	No	No		NusaCrowd carry-over	indolem_nerui	
294	NERGrit	Entity Sentiment	NER Grit dataset	https://github.con	-	ind	Rejang Lebong	Named Entiy Rec	Language	other	Other (other)	Unknown				12,518	2,521	2,398		sentences	17437	IndoNLP			2020		Free	-	No	No		NusaCrowd carry-over	nergrit	
295	NERP (IndoNLU Split)		NERP is a NER c LOC (name of lo or brand), EVT (r (name of food an	https://github.con	https://huggingfa	ind		Named Entiy Rec	Language	news articles	Creative Commo	Expert-generated				6,720	840	840		sentences	8400	IndoNLP	Investigating Bi-L	https://www.sema	2018		Free	-	No	No		NusaCrowd carry-over	nerp	
296	NERSkill.Id		NERSkill.Id is a r	https://data.menc		ind	Indonesia	Named Entity Re	Language	General Mixed	Creative Commo	Crawling	Manual (full)	2					418868	tokens	418868	Universitas PGRI	NERSkill.Id: Ann	https://www.scien	2023	ScienceDirect	Free		No	No	2024-04-03 18:0	Approved	nerskill_id	https://github.con
297	Netifier	Indonesian	Netifier dataset is	https://github.con		ind		Sentiment Analys	Language	Social media	Creative Commo	Crawling Expert-generated				6,995	0	778		tweets	7773	IndoNLP			2018		Free		No	No		NusaCrowd carry-over	netifier	
298	News En-Id MT		News En-Id is a r	https://github.con	-	eng ind		Machine Translat	Language	news articles	Creative Commo	Crawling Expert-translated				38,469	1,953	1,954		sentences	42376	IndoNLP	Benchmarking M	https://aclantholo	2021		Free	Pan Asia Networ	No	No		NusaCrowd carry-over	news_en_id	
299	NewsPH		Raw collection of	https://huggingfa	https://huggingfa	tgl fil	Phillipines	Language Model	Language	News	GNU General Pu	Crawling	None						2190000	instances	2190000	De La Salle Univ	Exploiting News	https://arxiv.org/a	2020	PRICAI	Free		No	No	2023-11-20 01:4	Approved	newsph	https://github.con
300	NIE Corpus of Spoken Singapore E	interviews rth_wind_and_th	This corpus cont	https://videoweb.		eng	Singapore	Automatic Speec	Language Speech	Conversational	Unknown (unkno	Expert-generated	None						N/A	utterances	0	National Institute	The NIE corpus c	https://fass.ubd.e	2001	No publication	Free		No	No	2023-12-13 17:1	Approved	nie_spoken_sg_e	https://github.con
301	NLLB Seed	Acehnese Banjarese Buginese	NLLB Seed is a s	https://tinyurl.con		eng ace bjn bug		Machine Translat	Language	Wikipedia	Creative Commo	Crawling Expert-translated				30,965	0	0		sentence pairs	30965	IndoNLP	No Language Lef	https://arxiv.org/p	2022		Free		No	No		NusaCrowd carry-over	nllb_seed	
302	NTREX-128	mya-fil mya-ind mya-khm mya-lao mya-zlm mya-tha mya-vie mya-hmv fil-mya fil-ind fil-khm fil-lao fil-zlm fil-tha fil-vie fil-hmv ind-mya ind-fil ind-khm ind-lao ind-zlm ind-tha ind-vie ind-hmv khm-mya khm-fil khm-ind khm-lao khm-zlm khm-tha khm-vie khm-hmv lao-mya lao-fil lao-ind lao-khm lao-zlm lao-tha lao-vie lao-hmv zlm-mya zlm-fil zlm-ind zlm-khm zlm-lao zlm-tha zlm-vie zlm-hmv tha-mya tha-fil tha-ind tha-khm tha-lao tha-zlm tha-vie tha-hmv vie-mya vie-fil vie-ind vie-khm vie-lao vie-zlm vie-tha vie-hmv hmv-mya hmv-fil hmv-ind hmv-khm hmv-lao hmv-zlm hmv-tha hmv-vie	NTREX-128, a d	https://github.con		mya fil ind khm lao zlm tha vie hmv	N/A	Machine Translat	Language	News	Creative Commo	Expert-generated	Manual (full)						1997	sentences	1997	Microsoft	NTREX-128 – Ne	https://aclantholo	2022	ACL Workshop	Free	English WMT19	No	No	2023-12-16 21:1	Approved	ntrex_128	https://github.con
303	NUS SMS Corpus	English Mandarin Chinese	This is a corpus c	https://github.con		eng cmn	Singapore	Language Model	Language	Conversational SMS	Unknown (unkno	Crowdsourced	None						55835 31465	utterances	87300	National Universi	Creating a Live, F	https://link.spring	2013	Language Resou	Free		No	No	2023-12-23 16:4	Approved	nus_sms_corpus	https://github.con
304	Nusantara Audiobook ASR	-pasentran-ke-ista turki salina	The dataset is a ' Nusantara Audio	https://nusantara	https://huggingfa	zlm	Malaysia	Automatic Speec	Language Speech	Books	Unknown (unkno	Machine-generat	Automatic						19.63 7.73 24.66	hours	52.02	Mesolitica	https://github.con		2024	No publication	Free	Nusantara Audio	No	No	2024-03-15 10:1	Approved	nusantara_audio	https://github.con
305	NusaParagraph Emot	btk bew bug jav mad mak min mui rej sun	NusaParagraph i	https://github.con		btk bew bug jav mad mak min mui rej sun		Emotion Recogni	Language	General	Creative Commo	Expert-generated				13,973	2,661	5,700		paragraphs	22334	IndoNLP	NusaWrites: Con	https://openrevie	2023		Free		No	No		NusaCrowd carry-over	nusa_paragraph_	
306	NusaParagraph Rhetoric	btk bew bug jav mad mak min mui rej sun	NusaParagraph i	https://github.con		btk bew bug jav mad mak min mui rej sun		Rhetoric Mode	Language	General	Creative Commo	Expert-generated				39,350	5,259	12,800		paragraphs	57409		NusaWrites: Con	https://openrevie	2023		Free					NusaCrowd carry-over	nusa_paragraph_	
307	NusaParagraph Topic	btk bew bug jav mad mak min mui rej sun	NusaParagraph i	https://github.con		btk bew bug jav mad mak min mui rej sun		Topic Modeling	Language	General	Creative Commo	Expert-generated				15,516	2,948	6,250		paragraphs	24714		NusaWrites: Con	https://openrevie	2023		Free					NusaCrowd carry-over	nusa_paragraph_	
308	NusaTranslation Emot	ind abs btk bew bhp jav mad mak min mui rej sun	NusaTranslation	https://github.con		ind abs btk bew bhp jav mad mak min mui rej sun		Emotion Recogni	Language	Hotel reviews Reviews Social media	Creative Commo	Expert-translated				21,800	3,131	8,800		sentences	33731		NusaWrites: Con	https://openrevie	2023		Free	IndoLEM Sentim				NusaCrowd carry-over	nusa_translation_	
309	NusaTranslation MT	ind abs btk bew bhp jav mad mak min mui rej sun	NusaTranslation	https://github.con		ind abs btk bew bhp jav mad mak min mui rej sun		Machine Translat	Language	Hotel reviews Reviews Social media	Creative Commo	Expert-translated				50,200	6,644	15,600		sentence pairs	72444		NusaWrites: Con	https://openrevie	2023		Free	IndoLEM Sentim				NusaCrowd carry-over	nusa_translation_	
310	NusaTranslation Senti	ind abs btk bew bhp jav mad mak min mui rej sun	NusaTranslation	https://github.con	https://github.con	ind abs btk bew bhp jav mad mak min mui rej sun		Sentiment Analys	Language	Hotel reviews Reviews Social media	Creative Commo	Expert-translated				24,810	3,503	10,400		sentences	38713		NusaWrites: Con	https://openrevie	2023		Free	IndoLEM Sentim				NusaCrowd carry-over	nusa_translation_	
311	NusaX MT	Acehnese Balinese Banjarese Buginese English Indonesian Javanese Madurese Minangkabau Ngaju Sundanese Toba Batak	The first-ever par	https://github.con	-	ind ace jav sun min bug bbc ban nij mad bjn eng		Machine Translat	Language	reviews	Creative Commo	Expert-translated				500	100	400		sentence pairs	1000		NusaX: Multilingu	https://arxiv.org/a	2022		Free	SmSA				NusaCrowd carry-over	nusax_mt	
312	NusaX Sentiment	Acehnese Balinese Banjarese Buginese English Indonesian Javanese Madurese Minangkabau Ngaju Sundanese Toba Batak	The first-ever par	https://github.con	-	ind ace jav sun min bug bbc ban nij mad bjn eng		Sentiment Analys	Language	reviews	Creative Commo	Expert-translated				500	100	400		sentences	1000		NusaX: Multilingu	https://arxiv.org/a	2022		Free	SmSA				NusaCrowd carry-over	nusax_senti	
313	OIL		The Online Indor	https://huggingfa	https://huggingfa	ind eng	Indonesia	Automatic Speec	Speech	Education	Creative Commo	Crawling Expert-generated Machine-generat	Manual (full)	1					98	examples	98	The Australian N The University of	Automated speec	https://aclantholo	2023	ACL	Free	YouTube	Yes	No	2023-12-30 03:1	Approved	oil	https://github.con
314	OJW		OJW is abbreviat	https://github.con		jav		Wordnet	Language	General	Unknown (unkno	Unknown				5,038	0	0		tokens	5038		Building the Old .	https://aclantholo	2020		Free					NusaCrowd carry-over	ojw	
315	Okapi m-ARC		m-ARC is a multi	http://nlp.uoregor	https://huggingfa	ind vie	N/A	Commonsense R	Language	Mixed	Creative Commo	Expert-translated Machine-translat	Automatic			N/A	N/A	N/A		examples	0	University of Ore	Okapi: Instructior	https://arxiv.org/a	2023	arXiv	Free	ARC	No	No	2024-02-27 14:0	Approved	okapi_m_arc	https://github.con
316	Okapi m-Hellaswag		m-Hellaswag is a	http://nlp.uoregor	https://huggingfa	ind vie	N/A	Question Answer	Language	Mixed	Creative Commo	Crowdsourced Machine-translat	Automatic			N/A	N/A	N/A		examples	0	University of Ore	Okapi: Instructior	https://arxiv.org/a	2023	arXiv	Free	Hellaswag	No	No	2024-02-27 14:2	Approved	okapi_m_hellasw	https://github.con
317	Okapi m-MMLU		m-MMLU is a mu	http://nlp.uoregor	https://huggingfa	ind vie	N/A	Commonsense R	Language	Mixed	Creative Commo	Expert-translated Machine-translat	Automatic			N/A	N/A	N/A		examples	0	University of Ore	Okapi: Instructior	https://arxiv.org/a	2023	arXiv	Free	MMLU	No	No	2024-02-27 14:1	Approved	okapi_m_mmlu	https://github.con
318	Okapi m-TruthfulQA		m-TruthfulQA is a	http://nlp.uoregor	https://huggingfa	ind vie	N/A	Question Answer	Language	Mixed	Creative Commo	Crowdsourced Machine-translat	Automatic			N/A	N/A	N/A		examples	0	University of Ore	Okapi: Instructior	https://arxiv.org/a	2023	arXiv	Free	TruthfulQA	No	No	2024-02-27 14:1	Approved	okapi_m_truthful	https://github.con
319	Onto4All		Onto4All is a sub	https://huggingfa	https://huggingfa	vie	Vietnam	Question Answer	Language	Mixed	Creative Commo	Crawling Crowdsourced Machine-translat	Manual (partial)						26300	examples	26300	Ontocord.ai	Onto4All: Enhanc	https://huggingfa	2024	No publication	Free		No	No	2024-03-16 11:3	Approved with notes	onto4all	https://github.con

No	Dataset name	Subset(s)	Dataset description	Dataset URL	HuggingFace URL	Dataset language(s)	Dataset collection region	Dataset task(s)	Dataset modality	Dataset domain(s)	Dataset license	Dataset annotation collection style	Dataset annotation validation style	Number of annotators per sample	Inter-annotator agreement	Train split data size	Validation split data size	Test split data size	No split data size	Data size unit	Total data size	Dataset provider	Dataset paper title	Dataset paper URL	Dataset or dataset paper publish year	Publication venue	Dataset access	Derived from...	Does the data have PII (personally identifiable information)?	Does the data have sensitive content/information?	Submission Date	Approval Status	Dataloader name	Github Issue URL
320	Open subtitles	Indonesian	OpenSubtitles is	https://opus.nlpl.e	https://huggingfa	ind eng		Machine Translat	Language	Transcribed audio Movies	GNU General Pu	Expert-translated				56,800	0	0		sentence pairs	56800		OpenSubtitles20	http://www.lrec-c	2016		Free					NusaCrowd carry-over	open_subtitles	
321	OpenAssistant	oasst1 oasst2	In an effort to der	https://github.com	https://huggingfa	tha vie ind	N/A	Chatbot	Language	Mixed	Apache license 2	Crowdsourced	Manual (partial)			N/A	N/A	N/A		dialogues	0	LAION	OpenAssistant C	https://arxiv.org/p	2023	NeurIPS	Free		No	No	2024-02-25 16:4	Approved	oasst2	https://github.com
322	OpenLID		This is an open d	https://github.com		ace ban bjn bug ceb ilo ind jav kac khm lao min lus mya pag shn sun tgl tha vie war zsm	N/A	Language Identif	Language	Multi-domain Wikipedia	GNU General Pu	Crawling	Manual (partial)						24223 15404 27667 7527 1002342 976648 1694230 65595 11365 60513 23529 31469 195262 452194 294618 21051 47458 1189616 734727 881145 282772 404380	sentences	8443735	University of Edi	An Open Dataset	https://aclantholo	2023	ACL	Free		No	No	2023-11-25 00:1	Approved	openlid	https://github.com
323	OpenMSD	tion (English -> No tion (Non English n (Non English -> on (English Non E (Non English No Coupling (English upling (Non Englis	Open-access Mu	https://github.com		ind vie tha zlm	N/A	Semantic Textual	Language	Academic Science	Creative Commo	Crawling Machine-generat	None			6M 11M 3M 208M 21M 1B 29M	2000 2000 2000 2000 400 11000 400	3000 6000 3000 1000 1000 7000 1000		documents	41800	Google	OpenMSD: Towa	https://arxiv.org/p	2023	Preprint	Free	The Unpaywall s	No	No	2023-12-26 08:1	Approved with notes	openmsd	https://github.com
324	OpenSLR	SLR80 (burmese) SLR35 (javanese) LR36 (sundanese LR44 (sundanese SLR42 (khmer) SLR41 (javanese)	This data set cor	https://www.open	https://huggingfa	mya jav sun khm	ar, Indonesia, Ca	Automatic Speec	Language Speech	General	Creative Commo	Crowdsourced	None			185000 220000 4210 2906 5822				utterances	420469	Google	Crowd-Sourced S	http://dx.doi.org/1	2018	SLTU	Free		No	No	2023-12-02 07:3	Approved	openslr	https://github.com
325	OpenSpeech Dataset V1 by Wang		The OpenSpeech	https://www.wang		tha	Thailand	Spoken Languag Language Model	Language Speech	Mixed	Creative Commo	Crowdsourced	Manual (full)	1077					8450	sentences	8450	Wang: data mark	N/A		2023	No publication	Free upon reque		No	No	2024-04-04 15:0	Approved with notes	openspeech_v1	https://github.com
326	OpenViVQA		OpenViVQA (Ope	https://huggingfa	https://huggingfa	vie	Vietnam	Visual Question A	Language Vision	General	MIT (mit)	Crawling Crowdsourced	Automatic & Man			9129	1070	1000		images	11199	Vietnam National University of Info	OpenViVQA: Tas	https://doi.org/10	2023	Information Fusic	Free		No	No	2023-12-29 03:0	Approved	openvivqa	https://github.com
327	Orchid Corpus		The ORCHID cor	https://github.com		tha	Thailand	POS Tagging	Language	Academic	Creative Commo	Expert-generated	Manual (full)						400000	tokens	400000	National Institute National Electron	ORCHID: Thai Pa	https://www.jstag	1998	Journal of the Ac	Free		No	No	2023-12-21 13:1	Approved	orchid_pos	https://github.com
328	OSCAR-2201		OSCAR or Open	https://huggingfa	https://huggingfa	ind zlm jav vie	N/A	Language Model	Language	General Mixed Multi-domain Web	BigScience Oper	Crawling	Automatic						N/A	tokens	0	Inria	Towards a Clean	https://aclantholo	2022	No publication	Free		Yes	Yes	2023-11-11 21:3	Approved	oscar_2201	https://github.com
329	PALITO Corpus		The PALITO Cor	https://github.com		tgl ceb ilo hil pag pam war bcl	N/A	Language Model Language Identif	Language	Books Religion	GNU Lesser Ger	Crawling	None						100,000 150,000	tokens	0	De La Salle Univ	Building Online C	https://aclantholo	2009	PACLIC	Free		No	No	2023-11-07 03:0	Approved	palito	https://github.com
330	Pangloss Collection	khm cog pcb sxm tpu jra thm bru kjg pkt nev oog hal tnu mya kac kkh lhu aem crw cje kjm mtq zng rgs tyr twh tpo viekhm	The Pangloss Cc	https://github.com		khm cog pcb sxm tpu jra thm bru kjg pkt nev oog hal tnu mya kac kkh lhu aem crw cje kjm mtq zng rgs tyr twh tpo vie	Cambodia Laos Myanmar Vietnam	Automatic Speec	Speech Language	Culture & heritage Dictionary	Creative Commo	Crowdsourced	Manual (partial)						Contact me for d	examples	0	Langues et Civilis Centre national d	Linguistic docum	https://hal.scienc	2001	Speech Commur	Free		No	No	2023-12-17 10:0	Approved	pangloss_collecti	https://github.com
331	PANL BPPT		Parallel Text Cor	https://huggingfa	-	eng ind		Machine Translat	Language	news articles	Unknown (unkno	Crawling Expert-generated				24,000	0	0		sentences	24000		Resource Report	https://aclantholo	2009		Free	-				NusaCrowd carry-over	id_panl_bppt	
332	ParaCotta		ParaCotta is a sy	https://drive.goog		ind eng		Paraphrasing	Language	News articles	Unknown (unkno	Machine-translat				6,000,000	0	0		sentence pairs	6000000		ParaCotta: Synth	https://aclantholo	2021		Free	ParaBank2, Wiki				NusaCrowd carry-over	paracotta_id	
333	Parallel Bible Corpus	Zyphe Chin Teduray Yakan	The Parallel Bibl	https://github.com		tiy yka zyp	N/A	Language Model Machine Translat	Language	Religion	Unknown (unkno	Crawling Expert-generated	None						7955 7957 8043	paragraphs	23955	Philipps Universi	Creating a massi	https://aclantholo	2014	LREC	Free		No	Yes	2023-12-29 17:3	Approved	parallel_bible_m	https://github.com
334	Parallel Corpus Dataset of Indones		This is a parallel This dataset is in	https://data.mend		ind zlm	engkulu, Indonesi	Machine Translat	Language	General Mixed	Creative Commo	Crawling Crowdsourced	Manual (full)						5261	sentence pairs	5261	Universitas Ahma	Statistical Machir	https://www.ijcao	2024	International Jou	Free		No	No	2024-04-03 17:0	Approved	ind_bengkulu_mt	https://github.com
335	Parallel: Indonesian - Lampung Ny		Parallel Indonesi	https://drive.goog	-	abl ind		Machine Translat	Language	general	Unknown (unkno	Crawling Expert-translated				3,000	0	0		sentences	3000		Effect of mono c	https://iopscience	2021		Free	Books				NusaCrowd carry-over	parallel_id_nyo	
336	ParaNames		ParaNames is a	https://github.com	https://huggingfa	bug ace ilo zlm tam bbc nia bjn nrm tet pag ceb vie tha eng dtp kjp bto lao min tgl por ban san shn war nod gor bcl cbk mnw jav mya mad krj khm lus btm cps pam abs hil ind	N/A	Named Entity Re Entity Linking Word-level Trans Word lists	Language	Wikipedia	Creative Commo	Crawling Crowdsourced	None			Contact me for d	N/A	N/A		instances	0	Brandeis Univers	ParaNames: A M	https://aclantholo	2022	SIGTYP	Free		No	No	2024-01-16 16:2	Approved with notes	paranames	https://github.com
337	PEx Conversations Dataset		PEx Conversatio The corpus cons	https://huggingfa	https://huggingfa	eng tgl	Philippines	Language Model	Language	Multi-domain Social media	Unknown (unkno	Crawling	None						Small Talk - 5K o Food & Drinks - 8 Health & Wellnes Body & Fitness - Home & Garden Style & Fashion - Travel & Leisure Visas & Immigrat	utterances	2418000	De La Salle Univ	Using Synthetic D	https://arxiv.org/a	2022	IALP	Free		No	No	2023-12-23 16:2	Approved	pex_conversatior	https://github.com
338	PFSA-ID	pfsa_id pfsa_id_med pfsa_id_test	PFSA-ID is an ar	https://github.com		ind	Indonesian	Named Entity Re Statement Taggir	Language	News articles Journalism	Creative Commo	Crawling	Automatic & Man		90.06	611 713 815	407 305 203	144		documents	3198	Universitas Atma	PFSA-ID: An Ann Extraction and at	https://doi.org/10. https://doi.org/10.		Global Knowledg Knowledge-Base	Free		Yes	No	2024-03-07 09:1	Approved	pfsa_id	https://github.com
339	Philippine Fake News Corpus		The Philippine Fa	https://github.com		eng	Philippines	Fact Checking	Language	News articles	Unknown (unkno	Crawling	Automatic			3081988	0	0		examples	3081988	Mapua University	Computing the Li	https://dl.acm.org	2019	9th International	Free		Yes	Yes	2023-11-01 14:4	Approved	ph_fake_news_c	https://github.com
340	PhoATIS		This is first public	https://github.com		vie	N/A	Slot Filling Intent Classificati	Language	airlines	Unknown (unkno	Expert-generated	Manual (full)			4478	500	893		utterances	5871	VinAI Research	Intent Detection a	https://arxiv.org/a	2021	Interspeech	Free	ATIS	No	No	2023-11-17 05:2	Approved	phoatis	https://github.com
341	PhoMT		We present a hig	https://github.com		vie eng	N/A	Machine Translat	Language	General	MIT (mit)	Crawling	Manual (full)			2977999	18719	19151		sentence pairs	3015869	VinAI Research	PhoMT: A High-Q	https://aclantholo	2021	EMNLP	Free		No	No	2023-11-21 19:2	Approved	phomt	https://github.com
342	PhoNER_COVID19	word_level syllabel-level	A named entity re	https://github.com		vie	N/A	Named Entity Re	Language	Health	Unknown (unkno	Crowdsourced	Manual (full)			5027	2000	3000		sentences	10027	VinAI Research	COVID-19 Name	https://aclantholo	2021	NAACL	Free		No	No	2023-11-21 19:4	Approved	pho_ner_covid	https://github.com
343	PhoST		PhoST is a high-	https://github.com		vie eng	N/A	Speech-to-Speec	Speech	Talk	Creative Commo	Crawling	Manual (full)			501.59	3.13	3.77		hours	6.9	VinAI Research	A High-Quality an	https://arxiv.org/a	2022	Interspeech	Free upon reque	TED-talk	No	No	2023-11-08 05:1	Approved	phost	https://github.com
344	POSP (IndoNLU Split)		POSP is an Indo	https://github.com	https://huggingfa	ind		POS Tagging	Language	news articles	Creative Commo	Expert-generated				6,720	840	840		sentences	8400		Investigating Bi-L	https://www.sema	2018		Free	-				NusaCrowd carry-over	posp	
345	PoSTagged Sundanese Monolingu		The dataset cons	https://dataverse.		sun		POS Tagging	Language	Journalistic blog	Creative Commo	Crawling Expert-generated				3,616	0	0		sentences	3616		PoSTagged Sunc	https://dataverse.	2015		Free					NusaCrowd carry-over	postag_su	
346	prachathai-67k		A large corpus of	https://github.com https://www.dropl		tha	Thailand	Text Classificatio Topic Classificati	Language	Multi-domain News articles News Politics Journalism	Apache license 2	Crawling Expert-generated	Manual (full)			70	10	20		documents	100	pyThaiNLP	WangchanBERTa	https://arxiv.org/p	2019	Preprint	Free		No	No	2024-03-10 16:5	Approved	prachathai_67k	https://github.com
347	PRDECT-ID		PRDECT-ID Data	https://data.mend		ind	Indonesia	Sentiment Analys Emotion Classific	Language	Reviews E-commerce	Creative Commo	Crawling	Manual (full)						5400	sentences	5400	Bina Nusantara U	PRDECT-ID: Indo	https://www.scien	2022	Data in Brief	Free		No	No	2023-12-31 17:2	Approved	prdect_id	https://github.com
348	ProSub	Common Annotations	ProSub is a colle The Common sul The Annotations	https://github.com		zsm ind jav tha vie mya	Malaysia Indonesia Thailand Vietnam Myanmar Japan	Word Sense Disa Word lists Semantic Role L Machine Translat	Language	Books Conversational Mixed Multi-domain Social media Talk Restaurant Folklore aily communicatio Entertainment Subtitles Family Friends YouTube videos Wordlist	Creative Commo	Expert-generated	None						7368 72834	forms	72834	Tokyo University	Pronoun substitu	https://www.anlp.	2023	Annual Meeting c	Free	Partially from TAI	No	No	2024-03-22 17:0	Approved with notes	prosub	https://github.com
349	QASiNa		Question Answer	https://github.com		ind	Indonesia	Question Answer	Language	Religion	MIT (mit)	Crawling Machine-generat	Manual (full)						500	examples	500	Bandung Institute MBZUAI	QASiNa: Religiou	https://arxiv.org/p	2023	Preprint	Free		No	No	2024-01-03 20:4	Approved	qasina	https://github.com
350	QED		QED - The QCRI It's developed by	https://opus.nlpl.e		eng vie tha mya jav ind tgl zlm ceb fil khm lao mad pam	Qatar	Machine Translat Language Model	Language	Entertainment Subtitles	Other (other)	Crawling	Automatic						N/A	sentence pairs	0	Qatar Computing	The AMARA Cor	https://aclantholo	2014	LREC	Free		No	No	2023-12-31 13:2	Approved with notes	qed	https://github.com
351	ROOTS_vi_ted		ROOTS_vi_ted is	https://huggingfa	https://huggingfa	vie	Vietnam	Language Model	Language	Web Talk	Creative Commo	Crawling	Automatic	1570		1570	N/A	N/A		examples	1570	HKUST	The BigScience R	https://arxiv.org/a	2022	NeurIPS	Free		No	No	2023-11-25 10:0	Approved	roots_vi_ted	https://github.com
352	SAInT	nalysis of Indones	SAInT is a datase	https://github.com		ind	Indonesian	Aspect Based Se	Language	Tourism	Creative Commo	Crawling	Automatic & Man						42969	comments	42969	Universitas Islam	Convolutional Ne	https://ieeexplore	2021	IEEE	Free		No	No	2024-05-12 23:2	Approved	saint	https://github.com
353	Saltik		Saltik is a datase	https://github.com		ind	Indonesia	Grammatical Erro	Language	General	GNU Affero Gene	Expert-generated	Manual (full)						3000	examples	3000	University of Indo	A Comparison Be	https://github.com	2023	ICAICTA	Free		No	No	2024-01-08 18:2	Approved	saltik	https://github.com
354	Sampiran	Indonesian	Sampiran is a da	https://github.com		ind		Poem Generation	Language	General	GNU Affero Gene	Crawling Expert-generated				7,879	0	0		sentences	7879		Sampiran: An Ind	https://github.com	2023		Free	online sources				NusaCrowd carry-over	sampiran	
355	Sarawak Malay		This is a Sarawal The data consist	https://github.com		zlm	Sarawak, Malaysia	Text-To-Speech S Automatic Speec	Speech	General	Creative Commo	Expert-generated	None						38	dialogues	38	Faculty of Compu	N/A		2019	No publication	Free		No	No	2024-02-11 10:3	Approved	sarawak_malay	https://github.com
356	scb-mt-en-th-2020		A Large English-T	https://github.com	https://huggingfa	tha eng	Thailand	Machine Translat	Language	Commentary General Hotel reviews Mixed Multi-domain News articles Reviews Social media News Wikidata	Creative Commo	Crawling Crowdsourced Expert-generated	Automatic & Man			801402	100173	100177		instances	1001752	AIResearch	A large English-T	https://link.spring	2020	Language Resou	Free	generated_review	No	No	2023-12-23 15:1	Approved	scb_mt_en_th	https://github.com
357	SCH		This is the first pu long-running Use	https://github.com		hnj	N/A	Language Model	Language	Websites	Creative Commo	Crawling	Automatic						858635	sentences	858635	Carnegie Mellon	A Hmong Corpus	https://aclantholo	2022	LREC	Free		No	Yes	2024-04-01 12:1	Approved	sch	https://github.com
358	SEA MADLAD	ace akb ban bbc bew bik btx ceb fil gor hil iba ilo ind jav kac khm kxd lao mad mak meo min mkn msa msi mya nij nut pag shn sun tet tha vie war	SEA MADLAD is Please note som MADLAD uses al	https://huggingfa	https://huggingfa	ace akb ban bbc bew btx ceb fil gor hil iba ilo ind jav kac khm kxd lao mad mak meo min mkn msi mya nij nut pag shn sun tet tha vie war	N/A	Language Model Language Identif	Language	Multi-domain Web	Creative Commo	Crawling	Automatic & Man						966 71 637 1.3K 2.7K 3.1K 1K 66.2K 901.5K 303 10.6K 7.6K 11.8K 38M 69.5K 2.6K 285.7K 46 216K 509 555 4.7K 1.5K 402 2.3M 1.9K 172.4K 183 67 1.6K 788 55K 40.4K 17.4M 55M 2.9K	documents	4527	Google Allen AI - UW	MADLAD-400: A	https://arxiv.org/a	2023	Preprint	Free	Common Crawl	Yes	Yes	2023-11-13 01:2	Approved	sea_madlad	https://github.com
359	SEA Wikipedia	ace ban bcl bjn bug cbk gor ilo ind jav map_bms khm lao mad min mnw mya nia pag pam shn sun tet tgl tha vie war	SEA Lang & Loca	https://huggingfa	https://huggingfa	ace ban bcl bjn bug cbk gor ilo ind jav map_bms khm lao mad min mnw mya nia pag pam shn sun tet tgl tha vie war	N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A	Language Model	Language	General Multi-domain Wikipedia Web	Creative Commo	Crowdsourced	Manual (partial)						12979 20611 14079 10503 9969 2242 5815254 15290 662443 15369 73080 11466 4897 1192 11839 225972 3271 348045 108819 1714 1108 8932 13662 61529 1464 159666 45121 1287910 1266204	documents	10214630	Independent	-		2023	No publication	Free	Wikipedia Huggir	Yes	Yes	2023-11-03 22:3	Approved	sea_wiki	https://github.com
360	Sea-bench	oning, General-ins oning, General-ins oning, General-ins easoning, Genera easoning, General easoning, Genera easoning, Genera asoning, General-	Sea-bench is a m	https://huggingfa	https://huggingfa	tgl khm vie tha lao mya ind zlm	SEA	Instruction Tuning	Language	Commentary Conversational Mixed Multi-domain Reviews Wikipedia News	Apache license 2	Expert-generated	Manual (full)						100 100 100 80 80 80 80 80	instances	700	Alibaba Group	SeaLLMs - Large	https://arxiv.org/p	2023	Preprint	Free	MT-bench	No	No	2023-12-06 03:2	Approved	seabench	https://github.com

No	Dataset name	Subset(s)	Dataset description	Dataset URL	HuggingFace URL	Dataset language(s)	Dataset collection region	Dataset task(s)	Dataset modality	Dataset domain(s)	Dataset license	Dataset annotation collection style	Dataset annotation validation style	Number of annotators per sample	Inter-annotator agreement	Train split data size	Validation split data size	Test split data size	No split data size	Data size unit	Total data size	Dataset provider	Dataset paper title	Dataset paper URL	Dataset or dataset paper publish year	Publication venue	Dataset access	Derived from...	Does the data have PII (personally identifiable information)?	Does the data have sensitive content/information?	Submission Date	Approval Status	Dataloader name	GitHub Issue URL
361	SEACrowd Instruct Multi-task Coll	eka Korpus - Tran puaNER - Transla STES-3K - Transla ews Filipino - Tra Speech Tiktok - Te pelling Normalizat cines-tweets - Trar kasi-bahasa - Trar ER-Tourism - Tran ese-MTRobustEv neSoda - Translat MABL - Translated MKQA - Translated Dataset from Indor ation with Alignm REX-128 - Transla MID-PAIR - Transl ofanity Dataset - ession Dataset - T y Tweet Corpus - Yolanda Tweets - T -ViCTSD - Transla -ViOCD - Transla Emotion Corpus (ealthQA - Transla i Sentiment Corpu ai Reviews - Tran XNLI - Translated oryCloze - Transla	The SEACrowd I	https://github.con	https://huggingfa	eng cmn ind mya vie tha fil tgl khg hmv hmf hnj lao zlm por tam yue khm jav abs ceb day xdy aoz	Southeast Asia	Instruction Tuning	Language	Mixed	Unknown (unkno	Expert-translated Machine-translat	Manual (partial)						332040	sentence pairs	332040	Massachusetts Ir	N/A		2024	No publication	Free	Bhinneka Korpus	No	No	2024-06-19 00:0	Approved with notes	seacrowd_instruc	https://github.con
362	SeaEval	Cross-MMLU Cross-LogiQA SG-Eval PH-Eval	SeaEval is a ben	https://github.con	https://huggingfa	ind vie zlm fil	Singapore Philipines	Commonsense R Question Answer	Language	Books Commentary Education General Health Legal Mixed Multi-domain Culture & heritage Ethics Software Academic Science	Creative Commo	Expert-generatec	Manual (full)						150 176 103 100	sentence pairs	529	Nanyang Techno Institute for Infocc	SeaEval for Multi	https://arxiv.org/p	2023	Preprint	Free	MMLU, LogiQA	No	No	2023-12-27 02:5	Approved	seaeval	https://github.con
363	Seahorse		SEAHORSE is a	https://storage.gc		vie	N/A	Summarization	Language	News articles WikiHow	Creative Commo	Expert-generatec Machine-generat	Manual (full)		0.82	60980	8968	18331		article-summary	88279	Google	Seahorse: A Mult	https://aclantholo	2023	EMNLP	Free	XSum, XL-Sum,	No	No	2023-12-22 14:5	Approved	seahorse	https://github.con
364	SEAME	train dev_sge dev_man	In Singapore and	https://github.con		cmn eng	Singapore Malaysia	Speech-to-Text T Automatic Speec Language Identif	Language Speech	Conversational	Unknown (unkno	Crowdsourced	Automatic & Man			101.13	11.42	N/A		hours	112.55	Nanyang Techno National Universi Universiti Sains M	SEAME:a manda	https://www.resea	2010	Interspeech	Free		No	No	2024-02-17 06:1	Approved	seame	https://github.con
365	Semeval STS Indo		SemEval is a ser	https://github.con		ind eng		Semantic Textual	Language	General	Unknown (unkno	Machine-translat				10,293	0	2,608		sentence pairs	12901				2019		Free					NusaCrowd carry-over	id_sts	
366	SentiBahasaRojak	movie product stock	This dataset cont	https://data.depo		zlm eng cmn	N/A	Sentiment Analys	Language	Reviews	Unknown (unkno	Crawling Machine-generat	Manual (partial)			566 722 521	63 81 58	70 90 114		examples	2285	National Central Academia Sinica	BRCCandSentiB	https://aclantholo	2022	COLING	Free		No	No	2024-02-27 10:3	Approved	senti_bahasa_roj	https://github.con
367	Sentiment-Annotated Taglish Produ		Sentiment-Annot	https://huggingfa	https://huggingfa	tgl eng	Philippines	Sentiment Analys Text Classificatio	Language	Hotel reviews Mixed Multi-domain Reviews Restaurant E-commerce	Creative Commo	Crawling	Manual (full)	3	0.83				10500	sentences	10500	Ateneo de Manila	N/A		2024	No publication	Free	N/A	No	No	2024-05-13 23:4	Approved	sentiment_taglist	https://github.con
368	Shopee Reviews Tagalog		The Shopee revi	https://huggingfa	https://huggingfa	tgl fil	Phillipines	Text Classificatio	Language	Reviews	Mozilla Public Lic	Crawling	None			10500	2250	2250		comments	15000	Pamantasan ng L	Enhancement to	https://uijrt.com/a	2023	UIJRT	Free		No	No	2023-11-20 02:1	Approved	shopee_reviews_	https://github.con
369	Singgalang		Singgalang is an	https://github.con	-	ind		Named Entiy Rec	Language	wikipedia	Unknown (unkno	Machine-generat				48,957	0	0		sentences	48957		Modified DBpedia	https://ieeexplore	2017		Free	Wikipedia, DBpe				NusaCrowd carry-over	singgalang	
370	SleukRith Set		SleukRith Set is t	https://github.con		khm	om Penh, Cambo Kandal, Cambodia em Reap, Camboc	Optical Characte	Language Vision	Culture & heritage	Unknown (unkno	Crowdsourced	Manual (full)			113206	N/A	90669		images	203875	University of La F Institute of Techn	A New Khmer Pa	https://dl.acm.org	2017	HIP	Free	SleukRith Set is	No	No	2023-12-20 03:2	Approved	sleukrith_ocr	https://github.con
371	SmSA		SmSA is a senter	https://github.con	https://huggingfa	ind		Sentiment Analys	Language	reviews	Creative Commo	Crawling Expert-generatec				11,000	1,260	500		sentences	12760		Improving Bi-LST	https://ieeexplore	2020		Free	Multiple Online S				NusaCrowd carry-over	smsa	
372	SNLI Indo		The SNLI Indo da Premise sentenc	https://data.menc		ind	N/A	Natural Languag	Language	Education	Creative Commo	Expert-translated	Manual (partial)	5		549365	9840	9822		sentence pairs	569027	Institut Teknologi	SNLI Indo: A recc	https://www.scien	2023	Data in Brief	Free	Stanford Natural	No	No	2023-12-29 13:1	Approved	snli_indo	https://github.con
373	Software Documentation Data Set f	EN-ID EN-TH EN-MS EN-VI	The data set orig	https://github.con		eng ind zlm tha vie	N/A	Machine Translat	Language	Software	Creative Commo	Expert-generatec	None			0 0 0 0	66 210 207 152	74 197 205 176		documents	1287	SAP	A Parallel Evalua	https://aclantholo	2020	Workshop on Asi	Free		No	No	2023-11-17 14:1	Approved	sap_wat	https://github.con
374	SPAMID-PAIR		SPAMID-PAIR is	https://data.menc		ind	Indonesia	Text Classificatio	Language	Social media	Creative Commo	Crawling	Manual (full)						72874	sentence pairs	72874	Gadjah Mada Un Universitas Kriste	SPAMID-PAIR: A	https://dx.doi.org/	2022	IJACSA	Free		No	No	2023-12-31 17:5	Approved	spamid_pair	https://github.con
375	SQuAD ID		Stanford Questio	https://github.con		ind		Question Genera	Language	General	Unknown (unkno	Machine-translat				102,657	11,407	10,567		documents	124631		Sequence-to-Sec	https://arxiv.org/p	2020		Free					NusaCrowd carry-over	squad_id	
376	SQuAD-ID-NLI	Indonesian	The SQuAD-ID-N	https://huggingfa	https://huggingfa	ind	N/A	Natural Languag	Language	General	Unknown (unkno	Machine-generat	Manual (partial)			236891	23749	23747		sentence pairs	284387	University of Indo Universitas Indor	Verification of QA		2024	NAACL	Free	SQuAD-ID	No	No	2024-03-21 23:1	Approved	squad_id_nli	https://github.con
377	SREDFM	SREDFM_vi	SREDFM is an a	https://github.con	https://huggingfa	vie	N/A	Relation Extractic	Language	Wikipedia Wikidata	Creative Commo	Machine-generat	Automatic						1.4 millions	relations	0	Babelscape	Proceedings of th	https://aclantholo	2023	ACL	Free		No	No	2023-11-13 04:2	Approved	sredfm	https://github.con
378	STB_EXT	auto_pos_stack uto_pos_multiviev en_ud_autopos gold_pos	We adopt the Un	https://github.con		eng	Singapore	Dependency Par POS Tagging	Language	General	MIT (mit)	Crawling Machine-generat	Manual (partial)			3050 3050 12543 3050	150 150 2002 N/A	150 150 2002 N/A		sentences	26297	UCSD SUTD	From Genesis to	https://dl.acm.org	2019	TALLIP	Free		No	No	2023-11-24 10:1	Approved	stb_ext	https://github.con
379	STIF-Indonesia		STIF-Indonesia i	https://github.con	-	ind		Style Transfer	Language	social media	MIT (mit)	Crawling Expert-generatec				1,922	214	363		sentences	2499		Semi-Supervised	https://arxiv.org/p	2020		Free	Twitter				NusaCrowd carry-over	stif_indonesia	
380	struct_amb_ind		This dataset cont	https://github.con		ind	West Java	Automatic Speec	Speech	Conversational	Unknown (unkno	Crowdsourced	Manual (full)			4160	320	320		utterances	4800	Bandung Institute University of Indo Japan Advanced	Speech Recognit	https://aclantholo	2023	EMNLP	Free		Yes	No	2023-12-27 09:1	Approved	struct_amb_ind	https://github.con
381	SU-ID ASR	Sundanese	This data set con	http://openslr.org		sun		Automatic Speec	Speech Language	General	Creative Commo	Expert-generatec				220,000	0	0		sentences	220000		Crowd-Sourced S	http://dx.doi.org/1	2018		Free					NusaCrowd carry-over	su_id_asr	
382	SU-ID TTS	Sundanese	This data set con	https://www.open		sun		Text-To-Speech S	Speech Language	General	Creative Commo	Expert-generatec				4,213	0	0		utterances	4213		A Step-by-Step P	http://dx.doi.org/1	2018		Free					NusaCrowd carry-over	su_id_tts	
383	Sundanese Twitter Dataset for Emc		Sunda Emotion c	https://github.con		sun		Emotion Classific	Language	Social media	Unknown (unkno	Crawling Expert-generatec				2,518	0	0		sentences	2518		Sundanese Twitt	https://ieeexplore	2020		Free					NusaCrowd carry-over	su_emot	
384	Sundanese-Indonesian Parallel Co		The dataset cons	https://dataverse.		ind sun		Machine Translat	Language	Wikipedia	Creative Commo	Expert-translated				3,616	0	0		sentence pairs	3616		Sundanese-Indor	https://dataverse.	2015		Free					NusaCrowd carry-over	parallel_su_id	
385	Tagalog Profanity Dataset		This dataset cont	https://huggingfa	https://huggingfa	tgl	Philippines	Abusive Languag Hate Speech Det Text Classificatio	Language	Social media	Unknown (unkno	Crawling	Manual (full)		75	11100	2780	0		tweets	13880	University of Mak	Context-Based P	https://papers.ssr	2023	Preprint	Free		No	Yes	2023-12-23 15:2	Approved	tgl_profanity	https://github.con
386	TalkBankDB CHILDES	Indonesian Javanese Manado Malay Tagalog Thai Yau	The Child Langu TalkBank is an in	https://github.con		ind jav xmm tgl tha jau	N/A	Automatic Speec	Language Vision Speech	Conversational	BSD 3-clause Cl	Expert-generatec	Manual (full)						N/A	utterances	0	Carnegie Mellon	The CHILDES Pr	https://direct.mit.e	2000	Computational Li	Free		Yes	Yes	2024-03-24 23:0	Approved	talkbankdb_child	https://github.con
387	TALPCo	Indonesian Korean Burmanese Malay Japanese Thailand Vietnamese	The TUFS Asian	https://github.con		ind jav mya tha vie zsm		Machine Translat	Language	General	Creative Commo	Unknown				0	0	0		sentences	0		TUFS Asian Lang	https://www.anlp.	2018		Free					NusaCrowd carry-over	talpco	
388	Tatabahasa		This test is a gen	https://github.con		zlm	Malaysia	Commonsense R	Language	Education	Unknown (unkno	Crawling	Manual (partial)						349	examples	349	Independent	NA		2023	NA	Free	https://tatabahas	No	No	2024-03-14 17:0	Approved	tatabahasa	https://github.con
389	Tatoeba	tatoeba.ind tatoeba.jav tatoeba.tgl tatoeba.tha tatoeba.vie	Tatoeba is a colle	https://tatoeba.or	https://huggingfa	ind vie tgl jav tha	N/A	Machine Translat	Language	Mixed	Other (other)	Crowdsourced	None			0	1100	5500		sentence pairs	6600	Association Tatoe	N/A	https://tatoeba.or	2006	No publication	Free		No	No	2023-10-27 13:4	Approved	tatoeba	https://github.con
390	Taxi1500 v2		Taxi1500 is a tex	Text: https://githu Label: https://gith		aaz zyp zsm yva xnn wrs tiy tha tgl tbl tbk smk ptu prf obo nlc nbq mya msm msk msb mqj mkn mbt mbs lex lbk ksw kmk kkl kje heg fil ebk dgc clu ceb cbk blz blw bkd bgs att atd atb amk alp agt abx agn	N/A	Text Classificatio	Language	General	Unknown (unkno	Crawling Crowdsourced	Automatic & Man			N/A	N/A	N/A		sentences	0	LMU München	Taxi1500: A Multi	https://arxiv.org/a	2023	arXiv	Free		No	No	2024-03-01 00:2	Approved	taxi1500_v2	https://github.con
391	TED En-Id MT		TED En-Id is a m	https://storage.gc	https://huggingfa	eng ind		Machine Translat	Language	mixed	Creative Commo	Crawling Expert-translated				87,406	2,677	3,179		sentences	93262		When and Why A	https://aclantholo	2021		Free	TED talk scripts				NusaCrowd carry-over	ted_en_id	
392	TermA		The TermA span-	https://github.con	https://huggingfa	ind		Keyword Extracti	Language	hotel reviews	Creative Commo	Expert-generatec				3,000	1,000	1,000		documents	5000		Aspect and Opini	https://ieeexplore	2019		Free	-				NusaCrowd carry-over	term_a	
393	Thai Constitution Dataset		A collection of sn	https://github.con		tha	N/A	Language Model Language Identif	Language	Legal	Apache license 2	Crawling	None						N/A	documents	0	Khon Kaen Unive	No paper	https://github.con	2020	No publication	Free		No	No	2023-11-21 03:5	Approved	thai_constitution	https://github.con
394	Thai Depression Dataset		The corpus is a c	https://zenodo.or		tha	Thailand	Text Classificatio	Language	Social media	Creative Commo	Crawling	Manual (full)			25077	3344	5015		sentences	33436	University of Hels Mahidol Universit Rootroo Ltd	Detecting Depres	https://aclantholo	2021	W-NUT	Free		No	No	2023-11-25 07:3	Approved	thai_depression	https://github.con
395	Thai Elderly Speech dataset	Healthcare Smart Home	The Thai Elderly	https://github.con https://www.wang		tha	Thailand	Automatic Speec	Language Speech	Health Smart home	Creative Commo	Crowdsourced	Manual (full)						2506 2650	sentences	5156	Data Wow VISAI AI			2022	No publication	Free		No	No	2024-03-30 18:4	Approved	thai_elderly_spee	https://github.con
396	Thai Gov v2 Corpus		The dataset cons	https://github.con		tha	Thailand	Language Model	Language	Banking Legal	Open Data Comr	Crawling	Automatic						Contact me (or b	documents	0	Independent VISTEC	Untitled		2023	No publication	Free		No	No	2024-01-13 22:3	Approved with notes	thaigov	https://github.con
397	Thai Handwritten Free Datasets by		The Thai Senten	https://www.wang		tha	Thailand	Optical Characte	Language Vision	Mixed	Creative Commo	Crowdsourced	Manual (full)	2026					10000	images	10000	Wang: data mark	N/A		2023	No publication	Free upon reque		No	No	2024-04-04 15:1	Approved with notes	thai_handwritten	https://github.con
398	THAI SER	StudioA StudioB Zoom	THAI SER datas The THAI SER c	https://github.con		tha	N/A	Speech Emotion	Speech	Commentary Conversational General	Creative Commo	Expert-generatec	None						18 62 20	recordings	100	Chulalongkorn U	-	https://airesearch	2021	No publication	Free		No	No	2023-11-17 13:5	Approved	thai_ser	https://github.con
399	Thai Toxicity Tweet Corpus		Thai Toxicity Twe	https://github.con	https://huggingfa	tha	Thailand	Text Classificatio	Language	Social media	Creative Commo	Crawling Expert-generatec	Manual (full)	3	0.78				3300	tweets	3300	Tokyo Metropolita	Annotation and C	https://www.ta-co	2018	Workshop on Tex	Free		Yes	Yes	2023-12-28 21:3	Approved	thai_toxicity_twee	https://github.con
400	Thai Winograd Schemas		This research inte	https://huggingfa	https://huggingfa	tha	Thailand	Commonsense R	Language	Multi-domain	Creative Commo	Expert-translated	None						285	sentences	285	New York Univers	Thai Winograd Sc	https://arxiv.org/p	2024	Preprint	Free		No	No	2024-07-03 14:3	Approved	thai_winograd	https://github.con
401	thai_exam		ThaiExam is a Th	https://huggingfa	https://huggingfa	tha	Thailand	Question Answer	Language	General School Exam	Apache license 2	Crawling	None			20+5+32+3+5 (in	N/A	112+95+135+67		examples	0	SCB 10X	Typhoon: Thai La	https://arxiv.org/a	2023	Preprint	Free		No	No	2024-06-26 00:5	Approved	thai_exam	https://github.con
402	Thai-dialect	khummuang korat pattani	Thai-dialect is a c	https://drive.goog		mfa nod tts	Thailand	Automatic Speec Speech-to-Text T Speech-to-Speec	Language Speech	E-commerce aily communicatio	Creative Commo	Crowdsourced Expert-generatec	Automatic & Man			khummuang: 237 korat: 38624 pattani: 25039	khummuang: 806 korat: 1080 pattani: 834	khummuang: 30 korat: 3852 pattani: 2793		utterances	0	Chulalongkorn U CMKL University	Thai Dialect Corp	https://www.isca-	2023	Interspeech	Free		No	No	2023-12-26 07:2	Approved		
403	Thai-joke-corpus		Thai jokes scrape	https://github.con		tha	Thailand	Language Model	Language	Multi-domain	GNU General Pu	Crawling Expert-generatec	Manual (partial)						449	instances	449	iAPP Technology	Survey on Thai N	https://github.con	2019	Language Resou	Free		No	No	2024-03-10 16:3	Approved	thai_joke_corpus	https://github.con
404	Thai-Lao Parallel Corpus		Thai-Lao Parallel	https://github.con		tha lao	Thailand, Laos	Machine Translat	Language	News	Creative Commo	Crawling	None						265	sentence pairs	265	Independent	-		2021	No publication	Free		No	No	2024-01-19 23:3	Approved	tha_lao_embassy	https://github.con
405	Thai-LM Dictionary		Thai lexicon cont	https://github.con		tha	Thailand	Abusive Languag Grammatical Errc Machine Translat Named Entity Re Toxic Wordlist Word lists Word Analogy Wordnet	Language	Multi-domain Wikipedia Dictionary wordnet Wordlist	Unknown (unkno	Expert-generatec	Manual (full)						N/A	phrases	0	Chulalongkorn U	A Comparative S	https://arxiv.org/p	2019	Asian Conferenc	Free		No	Yes	2024-03-10 14:4	Approved	thai_lm_dictionar	https://github.con
406	Thai-NNER		This work presen	https://github.con		tha	N/A	Named Entity Re	Language	General	Creative Commo	Crawling Crowdsourced	Automatic & Man			2935	979	980		documents	4894	Chulalongkorn U Vistec-AI	Thai Nested Nam	https://aclantholo	2022	ACL	Free		No	No	2023-11-17 14:3	Approved	thai_nner	https://github.con
407	thai-romanization		The Thai Romani	https://www.kagg		tha eng	Thailand	Word lists	Language	General	Creative Commo	Machine-generat	None						648,241	instances	0	Independent	N/A		2018	No publication	Free		No	No	2024-03-30 19:2	Approved	thai_romanizatio	https://github.con
408	Thai-Sum		We present Thai	https://github.con	https://huggingfa	tha	N/A	Summarization	Language	General	MIT (mit)	Crawling	None			358868	11099	11097		article-summary	381064	Beijing Institute o	Using Knowledge	https://www.resea	2020	No publication	Free upon reque		No	No	2023-11-19 14:5	Approved	thai_sum	https://github.con
409	thai-tnhc2-books		This dataset colle	https://www.arts.	https://huggingfa	tha	Thailand	Language Model	Language	Books	Creative Commo	Crawling	None						353	documents	353	pyThaiNLP	Thai TNHC2 Boo	https://doi.org/10.	2024	Zenodo	Free		No	No	2024-03-30 17:2	Approved	thai_tnhc2_books	https://github.con
410	ThaiSpoof		Thai language da	https://gofile-373		tha	Thailand	Hoax Detection Spoken Languag	Speech	Conversational Multi-domain	Creative Commo	Crowdsourced Machine-generat	Automatic & Man						143262 1575882	utterances	1719144	Universiti Brunei University of Con Japan Advanced NECTEC Thammasat Univ	ThaiSpoof: A Dat	https://ieeexplore	2023	ISAI-NLP	Free upon reque	Common Voice	No	No	2024-03-11 17:2	Approved	thai_spoof	https://github.con
411	TICO-19	TICO-19 id TICO-19 km TICO-19 ms TICO-19 my TICO-19 tl TICO-19 ta	Translation Initiat	https://tico-19.gitl		ind khm zlm mya tgl tam	N/A	Machine Translat	Language	Multi-domain	Creative Commo	Machine-generat	Automatic & Man			0 0 0 0 0 0	971 971 971 971 971 971	2100 2100 2100 2100 2100 2100		sentences	18426	multiple institutio	TICO-19: the Tra	https://aclantholo	2020	ACL Workshop	Free		No	No	2023-11-13 04:4	Approved	tico_19	https://github.con
412	TITML-IDN		TITML-IDN (Toky	https://huggingfa	https://huggingfa	ind		Automatic Speec	Speech Language	News articles	Other (other)	Expert-generatec				6,679	0	0		utterances	6679		A large vocabular	https://www.resea	2006		Free					NusaCrowd carry-over	titml_idn	
413	TLUnified-NER		This dataset cont	https://huggingfa	https://huggingfa	tgl	N/A	Named Entity Re	Language	General News	GNU General Pu	Expert-generatec	Manual (full)	3	0.81	6170	1543	3306		examples	11019	Independent	Developing a Nai	https://github.con	2023	IJCNLP-AACL	Free	TLUnified	No	No	2023-10-31 17:5	Approved	tlunified_ner	https://github.con
414	TMAD Malay Corpus		The Towards Mal	https://github.con		zlm	N/A	Word Sense Disa	Language	News	Unknown (unkno	Crowdsourced	Automatic & Man			4149	2075	2075		sentences	8299	Guangdong Univ	Towards Malay A	https://www.resea	2023	NLPCC	Free		No	No	2024-01-10 21:1	Approved	tmad_malay_cor	https://github.con
415	TotalDefMeme		This is a large-sc	Image: https://dri Annotations: http		eng	Singapore	Topic Classificati Stance Detection Optical Characte	Language Vision	Social media Memes National Policy	Unknown (unkno	Crowdsourced	Manual (full)	3					7200	images	7200	Singapore Univer	TotalDefMeme: A	https://arxiv.org/p	2023	ACM Multimedia	Free		No	No	2024-01-11 09:0	Approved	total_defense_m	https://github.con
416	Toxicity-200	Acehnese Banjarese Buginese Indonesian Javanese	Toxicity-200 is a	https://tinyurl.con		ind ace bjn bug jav		Toxic Wordlist	Language	General	Creative Commo	Expert-translated				226	0	0		phrases	226		No Language Lef	https://arxiv.org/p	2022							NusaCrowd carry-over	toxicity_200	
417	TrueVoice Intent		TrueVoice Intent	https://github.con		tha	Thailand	Intent Classificati	Language	Conversational Talk Customer Service	Apache license 2	Expert-generatec	Manual (full)			10303	0	2,608		dialogues	10303	pyThaiNLP	PyThaiNLP: Thai	https://aclantholo	2016	ACL Workshop	Free		No	No	2024-03-20 18:0	Approved	truevoice_intent	https://github.con
418	TSynC2 Corpus		TSynC2 is a Tha	https://github.con		tha	Thailand	Text-To-Speech S	Language Speech	News articles	Creative Commo	Expert-generatec	Manual (full)						7043	sentences	7043	NECTEC	TSynC-2: Thai S	https://lexitron.ne	2008	Technical Report	Free		No	No	2024-03-20 18:3	Approved	tsync2	https://github.con
419	TweetTaglish		A Dataset for Inv	https://github.con		tgl eng	N/A	Machine Translat	Language	Social media	Unknown (unkno	Crawling Expert-generatec	Manual (partial)						N/A	tweets	0	University of Illin	TweetTaglish: A D	https://aclantholo	2022	LREC	Free		No	No	2023-11-14 13:2	Approved	tweet_taglish	https://github.con
420	Twitter Corpus of Philippine Englis	TCOPE Public v1	This dataset con The publicly avai	https://osf.io/bjzp		eng	acolod, Philippine agupan, Philippine maguete, Philippir ral Santos, Philipp gan City, Philippin oilo City, Philippine azpi City, Philippi Manila, Philippines aga City, Philippine axas City, Philippin urigao, Philippine acloban, Philippine Vigan, Philippines	Dependency Par POS Tagging	Language	Social media	Creative Commo	Crawling	Automatic						N/A	tweets	0	The Chinese Uni	Broadening horiz	https://osf.io/k3qz	2023	English World-W	Free		No	No	2023-11-14 14:2	Approved	tcope	https://github.con
421	TydiQA	ary_task_indones primary_task_thai ndary_task_indon econdary_task_th	TyDi QA is a que	https://github.con	https://huggingfa	ind tha	N/A	Question Answer	Language	Mixed	Apache license 2	Crowdsourced	Manual (partial)			14952 11365	1805 2245	1809 2203		examples	34379	Google	TyDi QA: A Bencl	https://aclantholo	2020	TACL	Free		No	No	2023-12-11 18:5	Approved	tydiqa	https://github.con
422	TyDiQA-ID-NLI	Indonesian	The TyDIQA-ID-N	https://huggingfa	https://huggingfa	ind	N/A	Natural Languag	Language	Mixed	Apache license 2	Machine-generat	Manual (partial)			9695	1131	1711		sentence pairs	12537	University of Indo Universitas Indor	Verification of QA		2024	NAACL	Free	TyDIQA	No	No	2024-03-21 23:0	Approved	tydiqa_id_nli	https://github.con
423	Typhoon Yolanda Tweets		The dataset cont	https://github.con		fil	Philippines	Sentiment Analys	Language	Social media	Creative Commo	Crawling	Manual (full)	1		582	N/A	153		tweets	735	Bicol University	Sentiment Analys	https://arxiv.org/a	2018	Preprint	Free		No	No	2023-11-07 02:2	Approved	typhoon_yolanda	https://github.con
424	UCLA Phonetic Corpus		This dataset cont The dataset is pr	https://github.con		ace brv hil hni ilo khm mak mya pam	N/A	Automatic Speec	Speech	Talk	Creative Commo	Crowdsourced	Automatic & Man						39 45 51 22 90 70 49 116 57	utterances	539	Carnegie Mellon	Multilingual Phon	https://www.cs.cn	2021	ICASSP	Free	UCLA Phonetics	No	No	2023-12-14 08:1	Approved	ucla_phonetic	https://github.con
425	UD Javanese-CSUI		UD Javanese-CS	https://github.con		jav	Indonesia	Dependency Par	Language	General News	Creative Commo	Expert-generatec	Manual (full)			0	0	1000		sentences	1000	Universitas Indor	A Gold Standard		2023	No publication	Free	Javanese referer	No	No	2024-02-02 15:4	Approved	ud_jv_csui	https://github.con
426	UD_Indonesian-CSUI		The UD_Indones	https://github.con		ind		Dependency Par Machine Translat POS Tagging	Language	General News articles	MIT (mit)	Machine-generat Expert-generatec				656	0	374		sentences	1030		Tree Rotations fo	https://www.thesc	2020		Free					NusaCrowd carry-over	ud_id_csui	
427	UD_Indonesian-GSD		UD_Indonesian-(	https://github.con	-	ind		Dependency Par	Language	multi domain	Creative Commo	Expert-generatec				4,477	559	557		sentences	5593		Universal Depen	https://aclantholo	2013		Free	-				NusaCrowd carry-over	indolem_ud_id_g	
428	UD_Indonesian-PUD		An Indonesian de	https://github.con	-	ind		Dependency Par	Language	news articles wikipedia nonfiction	Creative Commo	Expert-generatec Expert-translated				1,000	0	0		sentences	1000		A Gold Standard	http://jaslli.org/file	2019		Free	-				NusaCrowd carry-over	ud_id_pud	

No	Dataset name	Subset(s)	Dataset description	Dataset URL	HuggingFace URL	Dataset language(s)	Dataset collection region	Dataset task(s)	Dataset modality	Dataset domain(s)	Dataset license	Dataset annotation collection style	Dataset annotation validation style	Number of annotators per sample	Inter-annotator agreement	Train split data size	Validation split data size	Test split data size	No split data size	Data size unit	Total data size	Dataset provider	Dataset paper title	Dataset paper URL	Dataset or dataset paper publish year	Publication venue	Dataset access	Derived from...	Does the data have PII (personally identifiable information)?	Does the data have sensitive content/information?	Submission Date	Approval Status	Dataloader name	GitHub Issue URL
429	UDHR	ind ilo mnw tet pam lus vie ccp min lao tgl hni ceb jav shn bcl hil sun ban zlm cnh kkh cfm ctd duu tdt tha bug mad mya khm war ace hnj blt hlt	he Universal Dec	https://huggingfa	https://huggingfa	ind ilo mnw tet pam lus vie ccp min lao tgl hni ceb jav shn bcl hil sun ban zlm cnh kkh cfm ctd duu tdt tha bug mad mya khm war ace hnj blt hlt	N/A	Language Model	Language	Education	Unknown (unkno	Expert-generatec	Manual (full)			1	0	0		documents	1	United Nations	United Nations. (	https://unicode.or	1998	No publication	Free	https://unicode.or	No	No	2023-11-16 02:3(	Approved	udhr	https://github.con
430	UDHR-LID	sun ace mad lao cfm hnj min zlm tha blt hni jav tdt cnh khm ban ind mya ccp duu tet kkh bug vie	The UDHR-LID d	https://huggingfa	https://huggingfa	sun ace mad lao cfm hnj min zlm tha blt hni jav tdt cnh khm ban ind mya ccp duu tet kkh bug vie	N/A	Language Identif	Language	Legal	Creative Commo	Expert-generatec	None						66 61 60 57 57 58 59 82 114 55 57 122 57 58 59 60 60 56 62 57 58 51 60 60	sentences	1546	Munich Center fo	GlotLID: Languag	https://arxiv.org/a	2023	EMNLP	Free	Universal Declar	No	No	2023-11-16 09:0(	Approved	udhr_lid	https://github.con
431	UIT-ViCoQA		UIT-ViCoQA is a	https://sites.goog		vie	Vietnam	Question Answer	Language	Health News articles	Other (other)	Expert-generatec	None			7000	1500	1500		instances	10000	Vietnam National University of Info	Conversational M	https://link.spring	2021	ICCCI	Free upon reque		No	No	2024-04-01 15:5	Approved	uit_vicoqa	https://github.con
432	UIT-ViCoV19QA	Answer 1 Answer 2 Answer 3 Answer 4	UIT-ViCoV19QA	https://github.con		vie	Vietnam	Question Answer	Language	Health	Unknown (unkno	Crawling Expert-generatec Machine-generat	Automatic & Man			3500 1390 542 272	500 209 79 39	500 201 79 39		examples	7350	University of Info	UIT-ViCoV19QA:	https://aclantholo	2022	PACLIC	Free		No	No	2023-12-30 22:2	Approved	uit_vicov19qa	https://github.con
433	UIT-ViCTSD	constructiveness toxicity	The UIT-ViCTSD	https://github.con	https://huggingfa	vie	Vietnam	Text Classificatio	Language	Education Health News Sports Science Business Cars Law World	Unknown (unkno	Crawling Expert-generatec	Manual (full)	3	0.58	7000 7000	2000 2000	1000 1000		comments	20000	Vietnam National University of Info	Constructive and	https://link.spring	2021	International Con	Free		No	Yes	2023-12-29 11:3	Approved	uit_victsd	https://github.con
434	UIT-ViIC		UIT-ViIC contains	https://drive.goog		vie	Vietnam	Image-to-Text Ge	Language Vision	Sports	Unknown (unkno	Expert-generatec	Manual (full)	5		2695	924	231		images	3850	Vietnam National University of Info	UIT-ViIC: A Datas	https://link.spring	2020	ICCCI	Free	MS-COCO	No	No	2023-12-26 00:3	Approved	uit_viic	https://github.con
435	UIT-ViOCD		The UIT-ViOCD c	https://huggingfa	https://huggingfa	vie	Vietnam	Text Classificatio	Language	E-commerce	Unknown (unkno	Expert-generatec	Manual (full)	3	0.87	4387	548	549		examples	5484	Vietnam National University of Info	Vietnamese Com	https://ebooks.ios	2021	New Trends in In	Free		No	No	2024-01-01 20:2	Approved	uit_viocd	https://github.con
436	UIT-ViON		UIT-ViON (Vietna	https://github.con		vie	Vietnam	Topic Classificati	Language	Education Health News Sports Science travel Business Entertainment Law World Technology Vehicle Life	Unknown (unkno	Crawling	Manual (partial)			208000	26000	26000		sentences	260000	University of Info	An Empirical Inve	https://ebooks.ios	2021	New Trends in In	Free		No	No	2023-12-29 14:1	Approved	uit_vion	https://github.con
437	UIT-ViQuAD		Vietnamese Que	https://sites.goog		vie	N/A	Question Answer	Language	Mixed	Unknown (unkno	Crowdsourced	Manual (full)						23,000	examples	0	University of Info	A Vietnamese Da	https://aclantholo	2020	COLING	Free upon reque		No	No	2024-03-15 21:5	Approved	uit_viquad	https://github.con
438	UIT-ViSD4SA		A Vietnamese da	https://github.con		vie	Vietnam	Aspect Based Se	Language	Reviews	Unknown (unkno	Expert-generatec	Manual (full)			0	0	0		comments		UIT Ho Chi Minh	Span Detection fo	https://aclantholo	2021	PACLIC	Free		No	No	2023-11-06 05:1	Approved	uit_visd4sa	https://github.con
439	UIT-ViSFD		UIT-ViSFD is the	https://github.con		vie	Vietnam	Aspect Based Se	Language	E-commerce	Unknown (unkno	Expert-generatec	None			7786	1112	2224		instances	11122	Vietnam National University of Info	SA2SL: From As	https://link.spring	2021	KSEM	Free		No	No	2024-04-01 15:4	Approved	uit_visfd	https://github.con
440	UIT-ViWikiQA		UIT-ViWikiQA is a	https://sites.goog		vie	Vietnam	Question Answer	Language	Wikipedia	Other (other)	Crowdsourced Expert-generatec Machine-generat	None			18,579 questions 4,101 passages 138 articles	2,285 questions 515 passages 18 articles	2,210 questions 493 passages 18 articles		instances	0	Vietnam National University of Info	Sentence Extract	https://arxiv.org/a https://dl.acm.org	2021	KSEM	Free upon reque	UIT-ViQuAD	No	No	2024-04-01 15:0	Approved	uit_viwikiqa	https://github.con
441	UKARA 1.0 Challenge		Ukara 1.0 Challe	https://simpan.ug		ind		Automatic Essay	Language	Education	Unknown (unkno	Expert-generatec				268	215	855		sentences	1338		UKARA 1.0 Chall	https://arxiv.org/a	2020		Free					NusaCrowd carry-over	ukara_challenge	
442	UniMorph	Indonesian Kodi Cebuano Hiligaynon Tagalog	The Universal Mo 5 Austronesian la	https://unimorph.		ind kod ceb hil tgl	nesia and Philippi	Morphological Inf	Language	Multi-domain	Creative Commo	Expert-generatec	None						27714 462 618 1256 2912	instances	32962	multiple institutio	UniMorph 4.0: Ur	https://aclantholo	2022	LREC	Free		No	No	2023-11-24 18:11	Approved	unimorph	https://github.con
443	Unimorph ID		The Universal Mo	https://github.con	-	ind		Morphological Inf	Language	books	Creative Commo	Crawling Expert-generatec				19,400	2,771	5,543		forms	27,714		SIGMORPHON 2 Generalization A	https://aclantholo	2021		Free	Unimorph				NusaCrowd carry-over	unimorph_id	
444	UniSent		UniSent is a univ	https://github.con		aaz abx ace agn agt ahk akb alj alp amk aoz atb atd att ban bbc bcl bgr bgs bgz bhp bkd bku blw blz bpr bps bru btd bth bto bts btx bug bvz bzq bzi cbk ceb cfm cgc clu cmo cnh cnw csy ctd czt dgc dtp due duo ebk fil gbi gor heg hil hnj hnn hvn iba ifa ifb ifk ifu ify ilo ind iry isd itv ium ivb ivv jav jra kac khm kix kje kmk kne kqe krj ksc ksw kxm lao lbk lew lex lhi lhu ljp lus mad mak mbb mbd mbf mbi mbs mbt mej mkn mnb mog mqj mqy mrw msb msk msm mta mtg wmm mtj mvp mwq mwv mya nbe nfa nia nij nlc ntx npy obo pag pam plw pmf	N/A	Sentiment Analys	Language	Religion Social media	Creative Commo	Machine-translat	Automatic & Man						N/A	instances	0	University of Cali	UniSent: Univers	https://aclantholo	2020	LREC	Free		No	No	2024-04-01 15:1	Approved	unisent	https://github.con
445	Universal Dependencies	id_gsd id_csui id_pud vi.vtb tl_trg tl_ugnayan	Universal Depen	http://hdl.handle.	https://huggingfa	ind vie tgl	N/A	POS Tagging	Language	Mixed	Apache license 2	Crowdsourced Expert-generatec	None			0	3753	0		examples	3753	UD Project	Universal Depen	https://aclantholo	2013	ACL	Free		No	No	2023-10-27 13:3	Approved	ud	https://github.con
446	University of Maryland Parallel Cor	Cebuano Bible Indonesian Bible /ietnamnese Bible	This dataset prov	http://users.umia		ceb ind vie	N/A	Machine Translat Language Model	Language	Religion	Unknown (unkno	Crawling Expert-generatec	None						31048 13468 7852	paragraphs	52368	University of Mar	The Bible as a Pa	https://link.spring	1999	Computers and th	Free		No	No	2023-12-26 21:1	Approved	bible_corpus_res	https://github.con
447	UP2.0	ind vie	Southeast Asian	https://github.con https://github.con		ind vie	N/A	Semantic Role La	Language	Mixed Multi-domain	Community Data	Expert-generatec Machine-generat	Automatic & Man			4477 1400	559 800	557 800		sentences	8593	IBM Research Kyndryl Vietnam National Apple	Universal Propos	https://aclantholo	2022	LREC	Free		No	No	2023-12-25 22:1	Approved	up2	https://github.con
448	Vi Pubmed		Vi Pubmed (or Vi	https://huggingfa	https://huggingfa	vie eng	N/A	Machine Translat	Language	Health	Other (other)	Machine-generat	None						20100000	documents	20100000	VietAI	Enriching Biomed	https://aclantholo	2022	EACL	Free	PubMed	No	No	2023-12-25 21:2	Approved	vi_pubmed	https://github.con
449	ViCon		ViCon comprises	https://www.ims.u		vie	N/A	Word Analogy	Language	Mixed	Creative Commo	Crowdsourced	Manual (full)						400 400 600	examples	1400	University of Stut	Introducing Two V	https://aclantholo	2018	NAACL	Free		No	No	2024-03-15 23:3	Approved	vicon	https://github.con
450	Vietnamese Hate and Offensive Sp		This dataset con	https://github.con		vie	Vietnam	Hate Speech Det	Language	Social media	MIT (mit)	Crawling	Manual (full)	6	0.86	8974	1122	1128		comments	11224	Vietnam National	ViHOS: Hate Spe	https://aclantholo	2023	EACL	Free	Vietnamese Hate	No	Yes	2023-12-23 10:2	Approved	vihos	https://github.con
451	Vietnamese Hate Speech Detection		The ViHSD datas	https://github.con		vie	Vietnam	Hate Speech Det	Language	Commentary Social media	Creative Commo	Crowdsourced	Manual (full)	4	0.5	24048	2672	6680		comments	33400	UIT Ho Chi Minh	A Large-Scale Da	https://link.spring	2021	International Con	Free		No	Yes	2024-01-02 10:5	Approved	uit_vihsd	https://github.con
452	Vietnamese Social Media Emotion		This dataset con Two rounds of m	https://drive.goog	https://huggingfa	vie	Vietnam	Emotion Classific	Language	Social media	Unknown (unkno	Crawling	Manual (full)		0.8	5548	686	693		comments	6927	Vietnam National	Emotion Recogni	https://arxiv.org/p	2019	PACLING	Free		No	No	2023-12-21 10:0	Approved	uit_vsmec	https://github.con
453	Vietnamese Students' Feedback Co		This corpus cons	https://drive.goog	https://huggingfa	vie	Vietnam	Sentiment Analys Topic Classificati	Language	Education Surveys	Unknown (unkno	Crowdsourced	Manual (full)			11426	1538	3166		sentences	16130	Vietnam National	UIT-VSFC: Vietna	https://ieeexplore	2018	Institute of Electr	Free		No	No	2023-12-21 09:3	Approved	uit_vsfc	https://github.con
454	ViGEText_17to23		The high-quality Objective was to General Educatio comprehensive a the years 2017 a Vietnamese stud the exact and un obtained from the	https://huggingfa	https://huggingfa	vie	Vietnam	Question Answer	Language	Education	Unknown (unkno	Crowdsourced	Manual (full)			35	765	3720		examples	4520	University of Info	Evaluating the Sy	https://dl.acm.org	2023	SOICT	Free		No	No	2024-03-16 00:1	Approved	vigetext	https://github.con
455	ViHealthQA		The ViHealthQA	https://huggingfa	https://huggingfa	vie	Vietnam	Question Answer	Language	Health	Unknown (unkno	Crawling Expert-generatec	Manual (full)			7009	993	2013		instances	10015	Vietnam National University of Info	SPBERTQA: A Tw	https://link.spring	2022	KSEM	Free		Yes	No	2024-01-01 20:4	Approved	vihealthqa	https://github.con
456	ViLexNorm		The ViLexNorm c	https://github.con		vie	Vietnam	Lexical Normaliz	Language	Social media	Creative Commo	Crowdsourced	Manual (full)			8373	1047	1047		sentence pairs	10467	UIT Ho Chi Minh	ViLexNorm: A Le	https://aclantholo	2024	EACL	Free		No	No	2024-03-19 22:5	Approved	vilexnorm	https://github.con
457	ViMMRC		ViMMRC, a chall	https://sites.goog		vie	N/A	Commonsense R	Language	Education	Unknown (unkno	Crowdsourced Expert-generatec	Manual (full)						2783	examples	2783	University of Info	Enhancing Lexica	https://ieeexplore	2020	IEEE Access	Free upon reque		No	No	2024-03-15 22:1	Approved	vimmrc	https://github.con
458	VIMQA		VIMQA, a new Vi	https://github.con		vie	N/A	Question Answer	Language	Mixed	Unknown (unkno	Expert-generatec	Manual (full)			8041	1003	1003		examples	10047	University of Scie	VIMQA: A Vietna	https://aclantholo	2022	LREC	Free upon reque		No	No	2024-03-15 21:5	Approved	vimqa	https://github.con
459	VinText		Vintext is a challe	https://github.con		vie	Vietnam	Optical Characte	Language Vision	General	GNU Affero Gene	Expert-generatec	Manual (full)	2	0.98	1200	300	500		images	2000	VinAI Research University of Ore Vietnam National University of Info University of Scie Stony Brook Univ	Dictionary-guidec	https://ieeexplore	2021	CVPR	Free		No	No	2023-12-23 20:4	Approved	vintext	https://github.con
460	ViSim-400		ViSim-400 is a da	https://www.ims.u		vie	N/A	Semantic Textual	Language	Mixed	Creative Commo	Crowdsourced	Manual (full)						400	examples	400	University of Stut	Introducing Two V	https://aclantholo	2018	NAACL	Free		No	No	2024-03-15 23:3	Approved	visim400	https://github.con
461	ViSoBERT		The ViSoBERT is	https://drive.goog	https://huggingfa	vie	Vietnam	Language Model	Language	Social media	Creative Commo	Crawling	None						1160	MB	1160	UIT Ho Chi Minh	ViSoBERT: A Pre	https://aclantholo	2023	EMNLP	Free		Yes	Yes	2024-01-02 01:0	Approved	visobert	https://github.con
462	ViSpamReviews		The dataset was	https://github.con		vie	Vietnam	Text Classificatio	Language	Reviews E-commerce	Creative Commo	Expert-generatec	Manual (full)	3		14306	1590	3974		comments	19870	UIT Ho Chi Minh	Detecting Spam	https://link.spring	2022	ACIIDS	Free		No	No	2024-01-02 12:0	Approved	vispamreviews	https://github.con
463	VISTEC-TP-TH-21		The largest socia	https://github.con		tha	N/A	Named Entity Re	Language	Social media	Creative Commo	Expert-generatec	Manual (partial)			39800	0	9900		sentences	49700	Vistec-AI	Handling Cross-	https://aclantholo	2021	ACL	Free		No	No	2023-11-17 15:1	Approved	vistec_tp_th_21	https://github.con
464	ViText2SQL		This is the first pu into Vietnamese.	https://github.con		vie	N/A	Text-to-SQL	SQL	Multi-domain	Unknown (unkno	Expert-generatec	None			6831	954	1906		sentences	9691	VinAI Research	A Pilot Study of T	https://aclantholo	2020	EMNLP Finding	Free	SPIDER	No	No	2023-11-13 04:0	Approved	vitext2sql	https://github.con
465	VIVOS		VIVOS is a free V	https://zenodo.or	https://huggingfa	vie	Vietnam	Automatic Speec	Language Speech	General	Creative Commo	Expert-generatec	None			11660	0	760		utterances	12420	Vietnam National	A non-expert Kal	https://aclantholo	2016	WLSI/OIAF4HLT	Free		No	No	2023-12-26 10:0	Approved	vivos	https://github.con
466	ViVQA		Vietnamese Visu	https://github.con		vie	Vietnam	Visual Question A	Language Vision	General	Unknown (unkno	Machine-generat	Automatic & Man			11999	0	3001		instances	15000	Vietnam National University of Info	ViVQA: Vietname	https://aclantholo	2021	ACL	Free	MS COCO and C	No	No	2023-12-29 02:3	Approved	vivqa	https://github.con
467	VlogQA		VlogQA is a Vietn	https://github.con		vie	Vietnam	Question Answer	Language	Food travel YouTube videos	Other (other)	Expert-generatec Machine-generat	Manual (partial)		0.44	8047	1017	1012		instances	10076	Vietnam National University of Info	VlogQA: Task, Da	https://aclantholo	2024	EACL	Free upon reque		No	No	2024-03-31 11:1	Approved	vlogqa	https://github.con
468	VLSP2016-NER		This dataset is co	https://vlsp.org.vn	https://huggingfa	vie	Vietnam	Named Entity Re	Language	Web	Creative Commo	Expert-generatec	Manual (full)			14918	0	2945		sentences	17863	VLSP	VLSP 2016 Shar	https://drive.goog	2016	Preprint	Free		Yes	Yes	2024-01-02 11:3	Approved	vlsp2016_ner	https://github.con
469	VLSP2016-SA		The SA-VLSP201	https://vlsp.org.vn		vie	Vietnam	Sentiment Analys	Language	Reviews	Creative Commo	Expert-generatec	Manual (full)	3		5100	0	1050		comments	6150	VLSP	VLSP SHARED T	https://vjs.ac.vn/i	2018	Journal of Compu	Free upon reque		Yes	Yes	2024-01-02 11:1	Approved	vlsp2016_sa	https://github.con
470	VLSP2020 MT		Parallel and mon	https://github.con		vie	Vietnam	Machine Translat	Language	Education Health Mixed News articles Wikipedia Entertainment Websites	Unknown (unkno	Crawling Expert-translated	Manual (partial)			4145592	2,227	789		sentence pairs	4146381	Karlsruhe Institut Vingroup Big Dat	Goals, Challenge	https://aclantholo	2020	VLSP	Free		No	No	2024-04-01 20:1	Approved	vlsp2020_mt_en	https://github.con
471	VLSP2020 RelEx		The dataset focu	https://docs.goog		vie	Vietnam	Relation Extracti	Language	News articles	Unknown (unkno	Expert-generatec	Manual (full)	2	0.7	606	250	300		documents	1156	Vietnam National VNU University o Viettel Group	Overview of VLS	https://aclantholo	2020	VLSP	Free upon reque		No	No	2024-04-01 20:2	Approved	vlsp2020_relex	https://github.con
472	VLSP2020 UDP		A gold universal	https://drive.goog	N/A	vie	Vietnam	Dependency Par	Language	Hotel reviews Mixed News articles News Restaurant	Unknown (unkno	Expert-generatec	Manual (full)	3	95.42	8152	N/A	1123		sentences	9275	VLSP	VLSP 2020 Shar	https://aclantholo	2020	VLSP	Free upon reque	VietTreebank, ot	No	No	2024-04-01 13:2	Approved	vlsp2020_udp	https://github.con

No	Dataset name	Subset(s)	Dataset description	Dataset URL	HuggingFace URL	Dataset language(s)	Dataset collection region	Dataset task(s)	Dataset modality	Dataset domain(s)	Dataset license	Dataset annotation collection style	Dataset annotation validation style
473	VLUE	UIT-ViQuAD 2.0 ViNLI VSMEC ViHOS NIIVTB POS	The VLUE bench	https://uitnlpgrou		vie	Vietnam	Text Classificatio POS Tagging Natural Languag Hate Speech Det	Language	Multi-domain	Creative Commo	Expert-generatec	None
474	VnDT: A Vietnamese dependency t	gold_standard predicted	VnDT is a Vietna Treebank consist	https://github.com		vie	N/A	Dependency Par	Language	General	Unknown (unkno	Machine-generat	Automatic
475	VoxLingua107		VoxLingua107 is	https://bark.phon	https://huggingfa	ceb ind jav zlm mya sun tha tgl vie war khm lao	N/A	Spoken Languag	Speech	YouTube videos	Creative Commo	Crawling	Automatic & Man
476	VoxLingua107	Indonesian Javanese Sundanese	VoxLingua107 is	http://bark.phon.i	https://huggingfa	ind sun jav		Spoken Languag	Speech Language	Social media	Creative Commo	Crawling Expert-generatec	
477	VSoLSCSum		The Vietnamese The dataset cont	https://github.com		vie	N/A	Summarization	Language	Mixed	Creative Commo	Crowdsourced	Manual (full)
478	WEATHub		WEATHub is a da	https://huggingfa	https://huggingfa	tha tgl vie cmn eng	N/A	Word lists	Language	General	Creative Commo	Expert-generatec	Manual (full)
479	WikiANN	id jv my tl th vi km su	WikiANN (someti	https://github.com	https://huggingfa	ind jav mya tgl tha vie khm sun	N/A	Named Entity Re	Language	Wikipedia	Apache license 2	Crawling Machine-generat	Manual (partial)
480	WikiAnn	ind jav bug min bjn sun ace tet ms map-bms	We developed a	https://drive.goog	https://huggingfa	ind jav bug min bjn sun ace tet zlm map-bms		Named Entiy Rec	Language	Wiki articles	Open Data Comr	Crawling	
481	WikiHow-GOSC	ind tha vie	This dataset con	https://github.com		ind tha vie	N/A	Goal-oriented Ge	Language	WikiHow	MIT (mit)	Crawling	Manual (full)
482	WikiLingua	English Indonesian	We introduce Wil a diverse set of t	https://drive.goog	https://huggingfa	ind eng		Cross-Lingual Ab	Language	WikiHow	Creative Commo	Crawling Expert-translated	
483	WikiMatrix	-de, ..., ceb-ru, ce	WikiMatrix is auto Wikipedia articles	https://github.com		ilo min jav sun ceb ind tgl vie	N/A	Machine Translat	Language	Wikipedia	Creative Commo	Machine-generat	Automatic
484	WikiText-TL-39		A benchmark Lar	https://s3.us-east	https://huggingfa	fil	N/A	Language Model	Language	Wikipedia	GNU General Pu	Crawling	None
485	WiLI-2018		WiLI-2018 is a W	https://zenodo.or	https://huggingfa	nrm jav min lao mya pag ind cbk tet tha ceb tgl bjn bcl vie	N/A	Language Identif	Language	Wikipedia	Open Database I	Crawling	None
486	Wisesight Thai Sentiment Corpus		This dataset com	https://github.com		tha	Thailand	Sentiment Analys Aspect Based Se	Language	Social media	Creative Commo	Crawling	Manual (full)
487	WIT		Wikipedia-based	https://github.com	https://huggingfa	ceb fil ind jav zlm mya tha vie war	N/A	Image-to-Text Ge Text-to-Image Ge	Language Vision	Wikipedia	Creative Commo	Crawling	Automatic & Man
488	Wongnai Reviews		Wongnai features Each write-up(re	https://github.com	https://huggingfa	tha	Thailand	Text Classificatio	Language	Reviews Restaurant	GNU Lesser Ger	Crowdsourced	Manual (full)
489	Word frequency distribution	Indonesian	Word frequency l	https://github.com		ind		Word lists	Language	General News articles Reviews Social media	Unknown (unkno	Crawling	
490	WReTe		WReTe is an ent	https://github.com	https://huggingfa	ind		Natural Languag	Language	other	Creative Commo	Crawling Expert-generatec	
491	X-FACT	Indonesian English	The largest publi	https://github.com		ind		Fact Checking	Language	News articles	MIT (mit)	Crawling Expert-generatec	
492	XCOPA	th id vi	XCOPA: A Multili	https://github.com	https://huggingfa	tha vie ind	N/A	Commonsense R	Language	General	Creative Commo	Expert-generatec	Manual (full)
493	XED	XED-VI	XED, a multilingu	https://github.com		vie	N/A	Sentiment Analys	Language	Web	Unknown (unkno	Expert-generatec Machine-translat	Manual (full)
494	XL-Sum	XL-Sum Burmese L-Sum Indonesia XL-Sum Thai -Sum Vietnamese	XL-Sum, a comp	https://github.com	https://huggingfa	mya ind tha vie	N/A	Abstractive Sumr	Language	News articles	Creative Commo	Crawling	Manual (partial)
495	XM3600	XM3600_fil XM3600_id XM3600_th XM3600_vi	Crossmodal-360 set of 3600 imag	https://google.git	https://huggingfa	fil ind tha vie	Southeast Asia	Image-to-Text Ge	Language Vision	Multi-domain	Creative Commo	Crawling Expert-generatec	Manual (full)
496	XNLI	vi th	XNLI is a subset	https://github.com	https://huggingfa	tha vie	N/A	Natural Languag	Language	Mixed	Creative Commo	Crowdsourced	Manual (full)
497	XPersona Id		XPersona is a op	https://storage.gc	-	ind		Dialogue System	Language	conversational	Creative Commo	Machine-translat Expert-translated	
498	XQuAD	xquad.th xquad.vi	XQuAD (Cross-li	https://github.com	https://huggingfa	tha vie	N/A	Question Answer	Language	Wikipedia	Creative Commo	Crowdsourced	None
499	XQuAD-R		XQuAD-R is a re	https://github.com		tha vie	N/A	Text Retrieval	Language	Wikipedia	Creative Commo	Expert-translated	None
500	xSID	English Indonesian	We introduce XS	https://bitbucket.c		ind eng		Spoken Languag	Speech Language	General	Creative Commo	Crawling Expert-translated	
501	XStoryCloze	id my	XStoryCloze con	https://huggingfa	https://huggingfa	ind mya	N/A	Commonsense R	Language	General	Creative Commo	Expert-generatec	Manual (full)
502	Yunshan-Cup-2020 Lao POS	lao	Lao POS dataset	https://github.com		lao	Laos	POS Tagging	Language	Web	Unknown (unkno	Expert-generatec	None

No	Number of annotators per sample	Inter-annotator agreement	Train split data size	Validation split data size	Test split data size	No split data size	Data size unit	Total data size	Dataset provider	Dataset paper title	Dataset paper URL	Dataset or dataset paper publish year	Publication venue	Dataset access	Derived from...	Does the data have PII (personally identifiable information)?	Does the data have sensitive content/information?	Submission Date	Approval Status	Dataloader name	GitHub Issue URL
473						N/A	instances	0	MBZUAI Vietnam National University of Info	VLUE: A New Be	https://arxiv.org/p	2024	Preprint	Free		No	No	2024-07-03 14:2	Approved with notes	vlue	https://github.com
474			8977 8977	200 200	1020 1020		sentences	20394	Vietnam National	From Treebank C	https://github.com	2014	NLDB	Free		No	No	2023-11-21 19:06	Approved	vndt	https://github.com
475			6628	1609	N/A		hours	8237	Independent	VoxLingua107: a	https://arxiv.org/a	2020	arXiv	Free		No	No	2024-02-13 11:10	Approved	voxlingua	https://github.com
476			157	0	0		hours	157		VoxLingua107: a	https://arxiv.org/p	2021		Free	YouTube				NusaCrowd carry-over	voxlingua	
477						141	documents	141	Japan Advanced	VSoLSCSum: Bu	https://aclantholo	2016	ALR	Free		No	No	2024-03-15 23:57	Approved	vsolscsum	https://github.com
478						150 175 77	examples	402	George Mason U	Global Voices, Lc	https://aclantholo	2023	EMNLP	Free		No	No	2023-11-20 02:09	Approved	weathub	https://github.com
479			786000	4980	10020		examples	801000	The University O Rensselaer Polyt	Massively Multilir	https://aclantholo	2019	ACL	Free	Wikipedia	No	No	2023-10-27 12:35	Approved	wikiann	https://github.com
480			254,240	0	0		Number of name	254240		Cross-lingual Nai	https://aclantholo	2017		Free					NusaCrowd carry-over	wikiann	
481			18913 5239 6255	0 0 0	2101 582 694		examples	33784	University of Pen	Goal-Oriented Sc	https://aclantholo	2021	INLG	Free		No	No	2024-03-13 15:44	Approved	wikihow_gosc	https://github.com
482			47,511	0	0		article-summary	47511		WikiLingua: A Ne	https://arxiv.org/p	2020		Free	WikiHow				NusaCrowd carry-over	wikilingua	
483						N/A	sentences	0	Meta	WikiMatrix: Minin	https://aclantholo	2021	EACL	Free	Wikipedia	No	No	2023-11-20 20:48	Approved	wikimatrix	https://github.com
484			39267089	8356898	8333288		tokens	55957275	De La Salle Univ	Evaluating Langu	https://arxiv.org/p	2019	Preprint	Free		No	Yes	2023-12-21 23:26	Approved	wikitext_tl_39	https://github.com
485			117500	0	117500		instances	235000	Independent	The WiLI benchm	https://arxiv.org/a	2018	Preprint	Free		No	Yes	2023-12-23 05:42	Approved	wili_2018	https://github.com
486			24066	N/A	2674		sentences	26740	Chulalongkorn U King Mongkut's In	No paper	https://github.com	2019	No publication	Free		No	No	2023-11-21 05:31	Approved	wisesight_thai_se	https://github.com
487			273344 37964 271133 20119 113190 17259 120592 515937 75167	1467 273 1865 130 728 107 793 3351 443	1388 353 1593 122 687 104 750 2696 334		examples	1461889	Google	WIT: Wikipedia-b	https://dl.acm.org	2021	SIGIR	Free		No	No	2023-12-19 09:49	Approved	wit	https://github.com
488			40000	0	6204		instances	46204	Wongnai	Wongnai Review	https://github.com	2019	No publication	Free		No	No	2023-12-09 11:18	Approved	wongnai_reviews	https://github.com
489			40,000	0	0		tokens	40000		Perbandingan dis	https://github.com	2013		Free	Kompas, Wikiped				NusaCrowd carry-over	freq_dist_id	
490			300	50	100		sentence pairs	450		Semi-supervised	https://www.resea	2018		Free	-				NusaCrowd carry-over	wrete	
491			2,231	297	448		evidences-links-c	2976		X-FACT: A New E	https://arxiv.org/p	2021		Free					NusaCrowd carry-over	x_fact	
492	1	0.911	0	0	109800		instances	109800	University of Can University of Mar	XCOPA: A Multili	https://aclantholo	2020	EMNLP	Free	COPA	No	No	2023-10-29 15:03	Approved	xcopa	https://github.com
493						956	examples	956	University of Hels	XED: A Multilingu	https://aclantholo	2018	COLING	Free		No	No	2024-03-15 10:59	Approved	xed	https://github.com
494			4569 38242 6616 32111	570 4780 826 4013	570 4780 826 4013		article-summary	101916	Bangladesh Univ	XL-Sum: Large-S	https://aclantholo	2021	ACL	Free	BBC	No	No	2023-11-08 04:56	Approved	xl_sum	https://github.com
495			0	0	3600 3600 3600 3600		examples	14400	Google	Crossmodal-360	https://aclantholo	2022	EMNLP	Free	Open Images	No	No	2023-11-15 20:53	Approved	xm3600	https://github.com
496			0	2380	0		examples	2380	Meta	XNLI: Evaluating	https://aclantholo	2018	ACL	Free	MultiNLI	No	No	2023-10-27 12:53	Approved	xnli	https://github.com
497			16,878	484	484		utterances	17846		XPersona: Evalu	https://aclantholo	2021		Free	PersonaChat				NusaCrowd carry-over	xpersona_id	
498			6896	1359	2357		examples	10612	Google HiTZ Center, Uni	On the Cross-ling	https://aclantholo	2020	ACL	Free	SQuAD 1.1	No	No	2023-10-27 13:03	Approved	xquad	https://github.com
499						1190 1190	instances	2380	Google Research	LAReQA: Langua	https://aclantholo	2020	EMNLP	Free	XQuAD	No	No	2024-04-01 06:22	Approved	xquadr	https://github.com
500			5,370,460	0	0		sentences	5370460		From Masked La	https://arxiv.org/p	2021		Free	Snips, Facebook				NusaCrowd carry-over	xsid	
501			360 360	0 0	1510 1510		sentences	3740	Meta	Few-shot Learnin	https://aclantholo	2022	EMNLP	Free	StoryCloze	No	No	2023-11-16 01:32	Approved	xstorycloze	https://github.com
502			6400	1600	4513		sentences	12513	Guangdong Univ	Yunshan Cup 20	https://arxiv.org/a	2022	Preprint	Free	lotenten corpus	No	No	2023-11-15 21:18	Approved	yunshan_cup_20	https://github.com

No	Dataset name	Dataset URL	Dataset paper URL	Submission Date
1	generated_reviews_enth	https://github.con	https://arxiv.org/p	2023-11-17 14:29
2	Dengue Filipino	https://huggingfa	https://ieeexplore	2023-11-20 01:58
3	Leipzig Corpora Collection	https://wortschat	http://www.lrec-c	2023-12-15 15:12
4	GlobalVoices	https://opus.nlpl.e	http://www.lrec-c	2023-12-19 18:39
5	DeepLontar	https://doi.org/10	https://www.natu	2023-12-23 17:47
6	Corpus Crawler	https://github.con		2023-12-23 20:50
7	OpenMSD	https://github.con	https://arxiv.org/p	2023-12-26 08:12
8	QED	https://opus.nlpl.e	https://aclantholo	2023-12-31 13:25
9	GNOME	https://opus.nlpl.e	https://aclantholo	2023-12-31 16:09
10	CCMatrix	https://opus.nlpl.e	https://aclantholo	2023-12-31 16:30
11	Lio and the Central Flores languages	https://archive.m	https://studentthe	2024-01-07 15:04
12	Thai Gov v2 Corpus	https://github.con		2024-01-13 22:38
13	ParaNames	https://github.con	https://aclantholo	2024-01-16 16:20
14	CLIRMatrix	https://github.con	https://aclantholo	2024-02-02 14:38
15	MM-Sum	https://drive.gooc	https://aclantholo	2024-02-17 10:34
16	BEST	https://github.con	http://pioneer.chu	2024-03-04 00:0
17	LEXiTRON	https://opend-por	https://www.sema	2024-03-10 16:12
18	multilingual-NLI-26lang-2mil7	https://huggingfa	https://www.camb	2024-03-16 01:47
19	Onto4All	https://huggingfa	https://huggingfa	2024-03-16 11:39
20	Amanatun wordlist	https://catalog.pa	https://www.ocse	2024-03-22 13:15
21	ProSub	https://github.con	https://www.anlp.	2024-03-22 17:02
22	InterBEST-2009	http://thailang.ne	https://ieeexplore	2024-03-23 13:20
23	BKD-Prosub	https://babyai-hu	https://www.virac	2024-03-29 11:12
24	AlloVera	https://github.con	https://aclantholo	2024-04-01 11:28
25	OpenSpeech Dataset V1 by Wang	https://www.wanc		2024-04-04 15:0
26	Thai Handwritten Free Datasets by Wang: Data Market	https://www.wanc		2024-04-04 15:10
27	SEACrowd Instruct Multi-task Collection	https://github.con		2024-06-19 00:04
28	Philippine Language Database (PLD)	Gitlab link not ye	https://aclantholo	2024-07-03 14:11
29	VLUE	https://uitnlpgrou	https://arxiv.org/p	2024-07-03 14:23

No	Dataset name	Dataset URL	Dataset paper URL	Submission Date	Notes
1	Belebele	https://github.con	https://github.con	2023-11-01 10:0	Duplicated #113
2	Dakshina	https://github.con	https://aclantholo	2023-11-21 00:2	**This Tamil is not from SEA**
3	SLR65	https://www.open	https://aclantholo	2023-11-30 00:4	**this Tamil dataset is not specifically collected from SEA region**
4	SLR80	https://www.open	https://aclantholo	2023-11-30 00:5	**Based on discussion with Holy, openSLR is considered as one single dataset. So, submissions ID 272 and 273 are put together as ID 278**
5	SLR42	https://www.open	https://www.isca-	2023-11-30 00:5	**Based on discussion with Holy, openSLR is considered as one single dataset. So, submissions ID 272 and 273 are put together as ID 278**
6	BLOOM LM	https://huggingfa	https://huggingfa	2023-12-03 21:1	Duplicated with #277
7	SEA Wikisource	https://huggingfa		2023-12-19 02:1	**It is another Wikipedia dump, with no peer-reviewed publication. We had already SEA Wiki, a Wikipedia dump dataset (submission #127) Another submission about Wikipedia dump is submission #308**
8	VISTEC-TP-TH-21	https://github.con	https://aclantholo	2023-12-21 13:0	duplicate to submission #237
9	Thai-Alpaca	https://huggingfa		2023-12-22 16:4	**The dataset uses gibberish Thai language—possibly due to bad results from Google Cloud Translation, which is used to translate the data from English to Thai. Two Thai researchers confirmed the data's unusability.**
10	Thai Databricks Dolly	https://huggingfa		2023-12-22 16:5	**The dataset uses gibberish Thai language—possibly due to bad results from Google Cloud Translation, which is used to translate the data from English to Thai. Two Thai researchers confirmed the data's unusability.**
11	Thai HH-RLHF	https://huggingfa		2023-12-22 16:5	**The dataset uses gibberish Thai language—possibly due to bad results from Google Cloud Translation, which is used to translate the data from English to Thai. Two Thai researchers confirmed the data's unusability.**
12	Thai GPTeacher	https://huggingfa		2023-12-22 17:0	**The dataset uses gibberish Thai language—possibly due to bad results from Google Cloud Translation, which is used to translate the data from English to Thai. Two Thai researchers confirmed the data's unusability.**
13	Singlish treebank	https://github.con	https://aclantholo	2023-12-23 17:0	**This dataset is covered in the latest release (STB_EXT, submission #323)**
14	Singlish treebank v2 (STB-EXT)	https://github.con	https://dl.acm.org	2023-12-23 17:1	duplicate to submission #258
15	eBible	https://github.con		2023-12-23 20:4	**This doesn't seem like a bible corpus, probably need to contact the submitter > Update 10 Jan, it is a bible corpus, the script crawl it in some ways I guess, no publication, how to proceed? > Update 23 Jan by Holy: Joel (the contributor) contacted me and said he had submitted a better version of eBible datasheet (no. 345). i'll make that one in review and reject this one.**
16	M3IT-80	https://huggingfa	https://arxiv.org/p	2023-12-25 23:3	duplicate with M3IT
17	JWSign	https://github.con	https://arxiv.org/a	2023-12-26 07:3	**The github page mentioned that the data should have been released on around November - December 2023 Per 1 April 2024, data is not available (NOT free upon request). Notes: As complementary information, JW300 dataset was pulled out from public access due to license problem**
18	Leipzig Corpora Collection	https://corpora.ur	https://link.spring	2023-12-27 23:2	duplicate with 289
19	MuMiN	https://data.bris.a	https://dl.acm.org	2023-12-29 16:2	**31/05/2024: All data cannot be decoded after unzipping. To compile the dataset a Twitter API key is required, which has become more difficult these days. However, the submission includes a link to a website where a full raw version of the dataset can be downloaded Among the languages listed, it is not clear if these Tweets were generated in Southeast Asia The other issue is that the non-English data is machine translated, so this data is not generated in SEA - Added the total number of Tweets, although the SEA language tweets will only be a very small percentage of the total - License is unclear, the submitted license is the one stated in the associated publication and in the readme file that downloads with the data, but the dataset website link states another license**
20	MuMiN	https://data.bris.a	https://dl.acm.org	2023-12-29 16:3	duplicate with submission #361
21	TED2020	https://opus.nlpl.(	https://arxiv.org/a	2024-01-01 18:0	**Invalid link**
22	CLICK-ID	https://data.menc	https://www.scier	2024-01-04 16:3	update data size (only the headlines) ---- holy: duplicate #47. removed its dataloader issue.
23	Multilingual Open Text (MOT)	https://github.con	https://aclantholo	2024-01-16 13:1	duplicate with no.400
24	Mesolithica muftiwp.gov.my scrape	https://huggingfa		2024-01-19 09:2	duplicate with no.411
25	Malaysia-AI government websites scrape	https://huggingfa		2024-01-21 05:0	**We decided that the crawled data from mesolitica https://huggingface.co/datasets/mesolitica/crawl-my-website Since there are 3 datasets pointing to these URL, we will just accept one of the three**
26	Malaysia AI GovDocs scrape	https://huggingfa		2024-01-21 05:1	**We decided that the crawled data from mesolitica https://huggingface.co/datasets/mesolitica/crawl-my-website Since there are 3 datasets pointing to these URL, we will just accept one of the three**
27	Aishell	https://www.aishe	https://arxiv.org/a	2024-02-11 11:3	**It is not indigenous language in SEA, while the data was not collected in SEA region**
28	THCHS-30	http://www.opens	https://arxiv.org/a	2024-02-11 17:5	**It is not indigenous language in SEA, while the data was not collected in SEA region**
29	PFSA-ID-MED	https://github.con	https://doi.org/10	2024-03-07 09:2	**Merge this on into #462**
30	PFSA-ID-TEST	https://github.con	https://doi.org/10	2024-03-07 09:3	**Merge this on into #462**
31	Thai Wikipedia Dump	https://github.con	https://arxiv.org/p	2024-03-10 16:2	**There is no wikipedia dump data provided in the repo, nonetheless there is the cleaned pretraining data of WangchanBERT: https://github.com/vistec-AI/thai2transformers/releases/tag/att-v1.0**
32	tamilmixsentiment	https://huggingfa	https://aclantholo	2024-03-15 12:0	**This Tamil is not from SEA region.**
33	tamil-alpaca	https://huggingfa	https://arxiv.org/a	2024-03-15 12:1	**This Tamil is not from SEA region.**
34	ULCA-ASR	https://github.con	https://bhashini.g	2024-03-15 12:3	**This Tamil is not from SEA region.**
35	IISc-MILE Tamil ASR Corpus	https://openslr.or	https://arxiv.org/a	2024-03-15 12:3	**This Tamil is not from SEA region.**
36	HopeEDI	https://github.con	https://aclantholo	2024-03-15 12:5	**This Tamil is not from SEA region.**
37	MinangNLP MT - Bilingual dictionary	https://github.con	https://aclantholo	2024-03-15 20:2	**MinangNLP MT has been carried-over from NusaCrowd**
38	CI-AVSR	https://github.con	https://aclantholo	2024-03-15 21:3	**This is Cantonese HK dataset.**
39	UIT-ViNewsQA	https://sites.goog	https://dl.acm.org	2024-03-15 22:0	**no clear information in the webpage how to obtain the dataset. consider to submit this data to private datasheet submission**
40	UIT-VSFC	https://drive.goog	https://ieeexplore	2024-03-15 22:1	duplicated with #304
41	OASST2	https://github.con	https://drive.goog	2024-03-16 11:1	duplicated #448? please check

No	Dataset name	Dataset URL	Dataset paper URL	Submission Date	Notes
42	multi-figqa	https://github.con		2024-03-18 18:1	duplicated with #225
43	Burmese speech dataset	https://www.open	https://aclantholo	2024-03-20 15:3	duplicated with #278
44	thai-policy-statements	https://github.con		2024-03-24 22:2	**This is NOT data for a summarization task.** **It is unannotated corpus. No clear metadata explaining the data.** **No peer-reviewed publication supports whether the data is collected using academic methodology.** **License is incorrect. Should be CC-0 instead of Apache 2.0**
45	Thai WIKI QA	https://aiforthai.in		2024-03-24 23:0	**The link does not provide dataset. It is a service / an API.** **If you thought that the service can provide the data, please submit into private datasheet.**
46	UIT-ViNewsQA	https://sites.goog	https://arxiv.org/p	2024-04-01 14:4	duplicated with #492
47	Multilingual Lexical Simplification - Filipino	https://github.con	https://sites.goog	2024-04-02 15:2	**The author of this dataset not allows anyone to re-release this dataset to any open-webs or any sites.**
48	id-hatespeech-detection	https://github.con	https://ieeexplore	2024-04-02 17:3	Duplicate from #28
49	final_project	https://github.con	https://arxiv.org/a	2024-04-03 16:1	Duplicate of #128, submitter did not check for existing datasheet There are 2 datasets on sentiment analysis of AiryRooms review already This submission includes the link to the paper author's Github page rather than the IndoNLP page

Here are potential SEA datasets to be included in SEACrowd, but not yet submitted.			
If a dataset has been submitted through the form, it will be automatically removed from this list if the dataset name is identical.			
No	**Dataset Name**	**Dataset URL**	**Notes**
1	CanVEC	https://github.con	https://aclanthology.org/2020.lrec-1.507.pdf