### Massively Multilingual Corpus of Sentiment Datasets and Multifaceted Sentiment Classification Benchmark

Łukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, Tomasz Kajdanowicz

WUST, Brand24, PUT

#### INTRO

How to build a multilingual sentiment model?

- single multi-lingual model vs. dedicated monolingual models
- training vs. fine-tuning
- transfer learning between domains
- transfer learning between languages

#### **METHOD**

- comprehensive linguistic typology
- manual selection of 79 datasets from the pool of 350 published datasets
- benchmarking 11 models (training mode, domain language, data modality, knowledge transfer)

#### RESULTS

- multilingual corpus of 79 high-quality sentiment datasets in 27 languages
- multi-faceted benchmark
- open library for dataset access

#### DISCUSSION

- large single multilingual model can perform approximately equally well for all languages
- all benchmarked models performed better when fine-tuned rather than trained from scratch
- bigger is better, but fine-tuning helps smaller models to be competitive
- large variability of performance across data splits



# The most extensive open massively multilingual corpus of sentiment datasets 80 datasets

## 27 languages 11 models





Take a picture to visit HF dataset and download the paper, dataset, code examples, code snippets, and benchmark

Table 1: Summary of the corpus	Categories: N - news, R	- reviews, SM -	- social media, O - other

				<u> </u>								
	#datasets		category				#samples		mean			
		N	R	SM	0	NEG	NEU	POS	#words	#chars		
English	17	3	4	6	4	304,939	290,823	1,734,724	62	339		
Arabic	9	0	4	4	1	138,899	192,774	600,402	52	289		
Spanish	5	0	3	2	0	108,733	122,493	187,486	26	150		
Chinese	2	0	2	0	0	117,967	69,016	144,719	60	80		
German	6	0	1	5	0	104,667	100,071	111,149	26	171		
Polish	4	0	2	2	0	77,422	62,074	97,192	19	123		
French	3	0	1	2	0	84,187	43,245	83,199	28	159		
Japanese	1	0	1	0	0	83,982	41,979	83,819	61	101		
Czech	4	0	2	2	0	39,674	59,200	97,413	34	212		
Portuguese	4	0	0	4	0	56,827	55,165	45,842	11	63		
Slovenian	2	1	0	1	0	33,694	50,553	29,296	41	269		
Russian	2	0	0	2	0	31,770	48,106	31,054	11	70		
Croatian	2	1	0	1	0	19,757	19,470	38,367	17	116		
Serbian	3	0	2	1	0	25,089	32,283	18,996	44	269		
Thai	2	0	1	1	0	9,326	28,616	34,377	22	381		
Bulgarian	1	0	0	1	0	13,930	28,657	19,563	12	86		
Hungarian	1	0	0	1	0	8,974	17,621	30,087	11	83		
Slovak	1	0	0	1	0	14,431	12,842	29,350	13	98		
Albanian	1	0	0	1	0	6,889	14,757	22,638	13	91		
Swedish	1	0	0	1	0	16,266	13,342	11,738	14	94		
Bosnian	1	0	0	1	0	11,974	11,145	13,064	12	76		
Urdu	1	0	0	0	1	5,239	8,585	5,836	13	69		
Hindi	1	0	0	1	0	4,992	6,392	5,615	26	128		
Persian	1	0	1	0	0	1,602	5.091	6,832	21	104		
Italian	2	0	0	2	0	4,043	4,193	3,829	16	103		
Hebrew	1	0	0	1	0	2,279	243	6,097	22	110		
Latvian	1	0	0	1	0	1,378	2,618	1,794	20	138		

Table 2: Models included in the benchmark													
Model	#params	#langs	base	reference									
mT5	277 M	101	T5	Xue et al. 93									
LASER	52M	93	BiLSTM	Artetxe and Schwenk 5									
mBERT	177M	104	BERT	Devlin et al. [26]									
MPNet	278M	53	XLM-R	Reimers and Gurevych 64									
XLM-R-dist	278M	53	XLM-R	Reimers and Gurevych 64									
XLM-R	278M	100	XLM-R	Conneau et al. [22]									
LaBSE	470M	109	BERT	Feng et al. 30									
DistilmBERT	134M	104	BERT	Sanh et al. 73									
mUSE-dist	134M	53	DistilmBERT	Reimers and Gurevych 64									
mUSE-transformer	85M	16	transformer	Yang et al. 95									
mUSE-cnn	68M	16	CNN	Yang et al. 95									

	Linear Head																													
XLM-R	52	51	61	58	48			67	58	63	61	58	39	49	53	45	48	58	49	53	53	38	56	53	41	45	45	55	54	41
LaBSE	54	53	61		53		57	65		63	60		44	55	55	52	48	55	57	52	54	41	53	57	34	42	52	61	55	53
MPNet	54	55	62	64	51	61		66	55	65	62	58	41	55	55	50	48	55	58	52	49	45	53		29	41	62	62	54	43
XLM-R-dist	52	51		61	45		58	64	55	62	61	55	41	53	49	51	47	52	55	52	52	43	54		35	46	50	56	49	45
mT5	50	48		56	45		56	65	53	63	54	57	39	49	52	39	44		52	47	41	39	54	49	35	40	48	52	57	48
mBERT	46	44		53	40	55	49	56	50	49	45	49	36	42	48	44	39	29	47	47	51	37	47	50	30	48	41	54	66	34
DistilmBERT	44	42	54	50	39	55	45	56	46	50	40	41	35	41	46	40	39	40	45	47	49	36	49	50	26	37	29	54	69	28
mUSE-dist	50	50	59	58	48		54	63	55		53	52	42	50	53	47	46	47	59	50	50	37	51	57	31	38	41	57	52	43
LASER	48	46		52	50		50		54	57	52	52	39	46	46	45	44	44	50	50	48	42	47	52	28	37	43	56	47	38
mUSE-transformer	45	47	55	55	48	57	52		51	56	52	40	43	41	50	42	40	45	46	52	43	39	46	48	28	40	29	54	27	23
mUSE-cnn	44	45	53	52	44	54	51	57	52	53	51	42	41	42	46	43	38	46	47	47	43	36	49	48	33	48	32	52	27	23
	Bus	ŝ	10	ė,	'n	ŝ	ĥ	ė	ā.	÷.	ŝ,	Ċ	¥.	Ŕ	ż	ź	à	£	ġ	Ż	÷	ġ.	Ś	Ś	É	ź	ġ	ж.	ġ	Ż
	-													BiL	STN	4 H	ead													
XLM-R	57	57	66	66	62	63	61	70	64	68	63	62	42	50	61	53	48	63	59	58	62	41	65	53	41	50	60	50	58	46
LaBSE	54	54	61	61	58	58	57	65	60	62	57	53	44	53	57	51	46	58	57	53	57	45		53	37	46	53			51
MPNet	55	56	63	64	61		58	66	58	64	60		43	54	54	50	50	60	58	57	54	43		57	32	39	58		57	52
XLM-R-dist	54	54	62	63	56	58	57	67	59	63	61		42	54	54	54	46	56	52	50	50	42	53	55	43	42		57	54	53
mT5	55	55	65	66	63	63	60	68		65		58	42	52	57	52	44	61	54	50	58	39		50	36	52	54	52	57	53
mBERT	50	49			56		51	62	56	56	49	48	38	48	56	48	40	44	50	51	55	36	51	53	35	48	40	55	57	37
DistilmBERT	49	48	58	57	57	57	52	61	53	55	46	45	39	47	57	48	40	42	48	46	56	36	56	51	34	44	40	59	70	39
mUSE-dist	53	53	61	62	60		56	64		62	54	56	41	49	53	51	44	48	53	54	54	42	55	55	37	42	50	57	52	51
	Bug	셤	Ē	ŝ	Ĺ6	ġ	ĥ	÷	ä	ė	.É,	Ś	Ŕ	'n	ż	È	ŝ	₽	ġ	Ż	Ť	ġ	à	Ŕ	5	ż	ġ	л£	Ż	ż
	-24													Fir	ne-t	uni	na													
XLM-R	61	61	68	70	66	64	64	72	67	70	68	65	44	58	66	55	50	65	61	62	62	45	61	55	50	51	63	61	58	65
LaBSE	60	62	67	69	67	63	61	71	67	70	64	63	42	59	66	58	50	63	61	64	63	47	61		43	48	65	62	57	61
MPNet	60	61	67	69	66	63	63	70	65	67	62	61	43	60	61	55	49	61		65	61	51	63	62	36	51	64	61	63	58
XLM-R-dist	59	61	67	68	66	62	62	70	65	67	63	61	45	59	62	58	50	62	58	62	62	49		60	41	48	62	62	61	56
mT5	57	60	66	68	65	63	62	70	64	70	64		44	53	60	49	44	62		53		41	60	55	38	54	52	61	52	45
mBERT	55	56	64	66	63		58	66	58	63	60		43	54	62	57	42	51	55	55	59	40	54	49	44	46	56	56	66	33
DistilmBERT	54	56	63	65	63		57	63	57	62	57	57	41	48	61	54	41	54	51	55	61	38	54	53	43	51	53	58	65	36
mUSE-dist	54	55	63	64	64			63		63	55	57	43	51	58	49	43	53	56	55	58	43	54	54	35	46	47	61	54	47
	Duc	쇵	ile all	8	ä	ġ	듃	÷	a.	÷	. <u>\$</u> ,	Ś	ň	'n	ż	ĥ	ŝ	£	Ż	Ż	¥	ä	2S	ź	'n	ż	ţ,	÷	2	ż

Figure 4: Detailed results of models' comparison. Legend: lang - averaged by all languages, ds -averaged by dataset, ar - Arabic, bg - Bulgarian, bs - Bosnian, cs - Czech, de - German, en - English, es - Spanish, fa - Persian, fr - French, he - Hebrew, hi - Hindi, hr - Croatian, hu - Hungarian, it -Italian, ja - Japanese, lv - Latvian, pl - Polish, pt - Portuguese, ru - Russian, sk - Slovak, sl - Slovenian, sq - Albanian, sr - Serbian, sv - Swedish, th - Thai, ur - Urdu, zh - Chinese.



Figure 5: Transfer learning between data modalities