cases	doc_1		doc_2		decision	id
	authors	• Andrei-Marius Avram • Darius Catrina • Dumitru-Clementin Cercel • Mihai DascÄƒlu • Traian Rebedea • Vasile PÄƒiÅŸ • Dan TufiÅŸ	**authors**	• Andrei-Marius Avram • D. Catrina • Dumitru-Clementin Cercel • Mihai Dascalu • Traian Rebedea • V. Pais • Dan Tufis	DUPLICATES	147
	title	Distilling the Knowledge of Romanian BERTs Using Multiple Teachers	**title**	Distilling the Knowledge of Romanian BERTs Using Multiple Teachers		
	publication_date	2022-04-13 00:00:00	**publication_date**	2021-12-23 00:00:00		
	source	SupportedSources.INTERNET_ARCHIVE	**source**	SupportedSources.OPENALEX		
	journal		**journal**	arXiv (Cornell University)		
	volume		**volume**			
	doi		**doi**	10.48550/arxiv.2112.12650		
	urls	• https://web.archive.org/web/20220523200334/https://arxiv.org/pdf/2112.12650v3.pdf	**urls**	• https://openalex.org/W4226067207 • https://doi.org/10.48550/arxiv.2112.12650 • http://arxiv.org/pdf/2112.12650		
	id	id-960900110111723557	**id**	id-3114794533438077014		
	abstract	Running large-scale pre-trained language models in computationally constrained environments remains a challenging problem yet to be addressed, while transfer learning from these models has become prevalent in Natural Language Processing tasks. Several solutions, including knowledge distillation, network quantization, or network pruning have been previously proposed; however, these approaches focus mostly on the English language, thus widening the gap when considering low-resource languages. In this work, we introduce three light and fast versions of distilled BERT models for the Romanian language: Distil-BERT-base-ro, Distil-RoBERT-base, and DistilMulti-BERT-base-ro. The first two models resulted from the individual distillation of knowledge from two base versions of Romanian BERTs available in literature, while the last one was obtained by distilling their ensemble. To our knowledge, this is the first attempt to create publicly available Romanian distilled BERT models, which were thoroughly evaluated on five tasks: part-of-speech tagging, named entity recognition, sentiment analysis, semantic textual similarity, and dialect identification. Our experimental results argue that the three distilled models offer performance comparable to their teachers, while being twice as fast on a GPU and ~35% smaller. In addition, we further test the similarity between the predictions of our students versus their teachers by measuring their label and probability loyalty, together with regression loyalty - a new metric introduced in this work.	**abstract**				
	versions		**versions**			