cases	doc_1		doc_2		decision	id
	authors	- KarolÃna BeneÅ¡ovÃ¡ - Andrej Å vec - Marek Å uppa	**authors**	- KatarÃna BeneÅ¡ovÃ¡ - Andrej Å vec - Marek Å uppa	DUPLICATES	192
	title	Cost-effective Deployment of BERT Models in Serverless Environment	**title**	Cost-effective Deployment of BERT Models in Serverless Environment		
	publication_date	2021-03-19 00:00:00	**publication_date**	2021-04-19 00:00:00		
	source	SupportedSources.OPENALEX	**source**	SupportedSources.INTERNET_ARCHIVE		
	journal	arXiv (Cornell University)	**journal**			
	volume		**volume**			
	doi	10.48550/arxiv.2103.10673	**doi**			
	urls	- https://openalex.org/W4287263520 - https://doi.org/10.48550/arxiv.2103.10673 - http://arxiv.org/pdf/2103.10673	**urls**	- https://web.archive.org/web/20210421032539/https://arxiv.org/pdf/2103.10673v2.pdf		
	id	id2424561014681025576	**id**	id-5406197705243831488		
	abstract		**abstract**	In this study we demonstrate the viability of deploying BERT-style models to serverless environments in a production setting. Since the freely available pre-trained models are too large to be deployed in this way, we utilize knowledge distillation and fine-tune the models on proprietary datasets for two real-world tasks: sentiment analysis and semantic textual similarity. As a result, we obtain models that are tuned for a specific domain and deployable in serverless environments. The subsequent performance analysis shows that this solution results in latency levels acceptable for production use and that it is also a cost-effective approach for small-to-medium size deployments of BERT models, all without any infrastructure overhead.		
	versions		**versions**			