cases	doc_1		doc_2		decision	id
	authors	- KarolÃna BeneÅ¡ovÃ¡ - Andrej Å vec - Marek Å uppa	**authors**	- Katar'ina Benevsov'a - Andrej vSvec - Marek vSuppa	DUPLICATES	190
	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**	2021-03-19 00:00:00		
	source	SupportedSources.OPENALEX	**source**	SupportedSources.SEMANTIC_SCHOLAR		
	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://www.semanticscholar.org/paper/2101193a3ec4d8fe260c7505614e760a7235ecf5		
	id	id2424561014681025576	**id**	id8821586626506311802		
	abstract		**abstract**	In this study we demonstrate the viability of deploying BERT-style models to AWS Lambda in a production environment. 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 the serverless environment. The subsequent performance analysis shows that this solution does not only report latency levels acceptable for production use but that it is also a costeffective alternative to small-to-medium size deployments of BERT models, all without any infrastructure overhead.		
	versions		**versions**			