cases	doc_1		doc_2		decision	id
	authors	• Katar'ina Benevsov'a • Andrej vSvec • Marek vSuppa	**authors**	• KatarÃna BeneÅ¡ovÃ¡ • Andrej Å vec • Marek Å uppa	DUPLICATES	234
	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 07:45:17+00:00		
	source	SupportedSources.SEMANTIC_SCHOLAR	**source**	SupportedSources.ARXIV		
	journal		**journal**	None		
	volume		**volume**			
	doi		**doi**			
	urls	• https://www.semanticscholar.org/paper/2101193a3ec4d8fe260c7505614e760a7235ecf5	**urls**	• http://arxiv.org/pdf/2103.10673v2 • http://arxiv.org/abs/2103.10673v2 • http://arxiv.org/pdf/2103.10673v2		
	id	id8821586626506311802	**id**	id4327090977737325995		
	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.	**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**			