

cases	doc_1		doc_2		decision	id		
					DUPLICATES	259		
	authors	<ul style="list-style-type: none">Xiao GuoHengameh MirzaalianEkraam SabirAyush JaiswalWael AbdAlmageed	authors	<ul style="list-style-type: none">Xiao GuoHengameh MirzaalianEkraam SabirAyush JaiswalWael Abd-Almageed				
	title	CORD19STS: COVID-19 Semantic Textual Similarity Dataset	title	CORD19STS: COVID-19 Semantic Textual Similarity Dataset				
	publication_date	2020-07-05 00:00:00	publication_date	2020-07-05 22:23:37+00:00				
	source	SupportedSources.OPENALEX	source	SupportedSources.ARXIV				
	journal	arXiv (Cornell University)	journal	None				
	volume		volume					
	doi	10.48550/arxiv.2007.02461	doi					
	urls	<ul style="list-style-type: none">https://openalex.org/W3039643360https://doi.org/10.48550/arxiv.2007.02461http://arxiv.org/pdf/2007.02461	urls	<ul style="list-style-type: none">http://arxiv.org/pdf/2007.02461v2http://arxiv.org/abs/2007.02461v2http://arxiv.org/pdf/2007.02461v2				
	id	id-8614122339413125268	id	id-2163659931015474250				
	abstract		abstract	In order to combat the COVID-19 pandemic, society can benefit from various natural language processing applications, such as dialog medical diagnosis systems and information retrieval engines calibrated specifically for COVID-19. These applications rely on the ability to measure semantic textual similarity (STS), making STS a fundamental task that can benefit several downstream applications. However, existing STS datasets and models fail to translate their performance to a domain-specific environment such as COVID-19. To overcome this gap, we introduce CORD19STS dataset which includes 13,710 annotated sentence pairs collected from COVID-19 open research dataset (CORD-19) challenge. To be specific, we generated one million sentence pairs using different sampling strategies. We then used a finetuned BERT-like language model, which we call Sen-SCI-CORD19-BERT, to calculate the similarity scores between sentence pairs to provide a balanced dataset with respect to the different semantic similarity levels, which gives us a total of 32K sentence pairs. Each sentence pair was annotated by five Amazon Mechanical Turk (AMT) crowd workers, where the labels represent different semantic similarity levels between the sentence pairs (i.e. related, somewhat-related, and not-related). After employing a rigorous qualification tasks to verify collected annotations, our final CORD19STS dataset includes 13,710 sentence pairs.				
	versions		versions					