	doc_1		doc_2		decision	id
cases	**authors**	• Marco Di Giovanni • Marco Brambilla	**authors**	• Marco Di Giovanni • M. Brambilla	DUPLICATES	238
	title	Exploiting Twitter as Source of Large Corpora of Weakly Similar Pairs for Semantic Sentence Embeddings	**title**	Exploiting Twitter as Source of Large Corpora of Weakly Similar Pairs for Semantic Sentence Embeddings		
	publication_date	2021-10-05 00:00:00	**publication_date**	2021-10-05 00:00:00		
	source	SupportedSources.INTERNET_ARCHIVE	**source**	SupportedSources.SEMANTIC_SCHOLAR		
	journal		**journal**	ArXiv		
	volume		**volume**	abs/2110.02030		
	doi		**doi**	10.18653/v1/2021.emnlp-main.780		
	urls	• https://web.archive.org/web/20211007050147/https://arxiv.org/pdf/2110.02030v1.pdf	**urls**	• https://www.semanticscholar.org/paper/e50c73f3b7b4556f282ae1d9b41c18fad1979883		
	id	id9074049210844344459	**id**	id-1385644731994850416		
	abstract	Semantic sentence embeddings are usually supervisedly built minimizing distances between pairs of embeddings of sentences labelled as semantically similar by annotators. Since big labelled datasets are rare, in particular for non-English languages, and expensive, recent studies focus on unsupervised approaches that require not-paired input sentences. We instead propose a language-independent approach to build large datasets of pairs of informal texts weakly similar, without manual human effort, exploiting Twitter's intrinsic powerful signals of relatedness: replies and quotes of tweets. We use the collected pairs to train a Transformer model with triplet-like structures, and we test the generated embeddings on Twitter NLP similarity tasks (PIT and TURL) and STSb. We also introduce four new sentence ranking evaluation benchmarks of informal texts, carefully extracted from the initial collections of tweets, proving not only that our best model learns classical Semantic Textual Similarity, but also excels on tasks where pairs of sentences are not exact paraphrases. Ablation studies reveal how increasing the corpus size influences positively the results, even at 2M samples, suggesting that bigger collections of Tweets still do not contain redundant information about semantic similarities.	**abstract**	Semantic sentence embeddings are usually supervisedly built minimizing distances between pairs of embeddings of sentences labelled as semantically similar by annotators. Since big labelled datasets are rare, in particular for non-English languages, and expensive, recent studies focus on unsupervised approaches that require not-paired input sentences. We instead propose a language-independent approach to build large datasets of pairs of informal texts weakly similar, without manual human effort, exploiting Twitterâ€™s intrinsic powerful signals of relatedness: replies and quotes of tweets. We use the collected pairs to train a Transformer model with triplet-like structures, and we test the generated embeddings on Twitter NLP similarity tasks (PIT and TURL) and STSb. We also introduce four new sentence ranking evaluation benchmarks of informal texts, carefully extracted from the initial collections of tweets, proving not only that our best model learns classical Semantic Textual Similarity, but also excels on tasks where pairs of sentences are not exact paraphrases. Ablation studies reveal how increasing the corpus size influences positively the results, even at 2M samples, suggesting that bigger collections of Tweets still do not contain redundant information about semantic similarities. Code available at https://github.com/marco-digio/Twitter4SSE		
	versions		**versions**			