cases	doc_1		doc_2		decision	id
	authors	• Sido, J. • SejÃ¡k, M. • PraÅ¾Ã¡k, O. • KonopÃk, M. • Moravec, V.	**authors**	• Jakub Sido • Michal SejÃ¡k • OndÅ™ej PraÅ¾Ã¡k • Miloslav KonopÃk • VÃ¡clav Moravec	DUPLICATES	105
	title	Czech News Dataset for Semantic Textual Similarity	**title**	Czech News Dataset for Semantic Textual Similarity		
	publication_date	2022-10-26 00:00:00	**publication_date**	2021-08-19 14:20:17+00:00		
	source	SupportedSources.CROSSREF	**source**	SupportedSources.ARXIV		
	journal		**journal**	None		
	volume		**volume**			
	doi	10.21203/rs.3.rs-2130964/v1	**doi**			
	urls	• https://www.researchsquare.com/article/rs-2130964/v1 • https://www.researchsquare.com/article/rs-2130964/v1.html • http://dx.doi.org/10.21203/rs.3.rs-2130964/v1	**urls**	• http://arxiv.org/pdf/2108.08708v3 • http://arxiv.org/abs/2108.08708v3 • http://arxiv.org/pdf/2108.08708v3		
	id	id3897652436991852601	**id**	id-9049170441244168679		
	abstract		**abstract**	This paper describes a novel dataset consisting of sentences with semantic similarity annotations. The data originate from the journalistic domain in the Czech language. We describe the process of collecting and annotating the data in detail. The dataset contains 138,556 human annotations divided into train and test sets. In total, 485 journalism students participated in the creation process. To increase the reliability of the test set, we compute the annotation as an average of 9 individual annotations. We evaluate the quality of the dataset by measuring inter and intra annotation annotators' agreements. Beside agreement numbers, we provide detailed statistics of the collected dataset. We conclude our paper with a baseline experiment of building a system for predicting the semantic similarity of sentences. Due to the massive number of training annotations (116 956), the model can perform significantly better than an average annotator (0,92 versus 0,86 of Person's correlation coefficients).		
	versions		**versions**			