

cases	doc_1		doc_2		decision	id
					DUPLICATES	242
	authors	<ul style="list-style-type: none">Hamid BekamiriD. HainRoman Jurowetzki	authors	<ul style="list-style-type: none">Hamid BekamiriDaniel S. HainRoman Jurowetzki		
	title	PatentSBERTa: A Deep NLP based Hybrid Model for Patent Distance and Classification using Augmented SBERT	title	PatentSBERTa: A Deep NLP based Hybrid Model for Patent Distance and Classification using Augmented SBERT		
	publication_date	2021-03-22 00:00:00	publication_date	2021-03-22 15:23:19+00:00		
	source	SupportedSources.SEMANTIC_SCHOLAR	source	SupportedSources.ARXIV		
	journal		journal	None		
	volume		volume			
	doi		doi			
	urls	<ul style="list-style-type: none">https://www.semanticscholar.org/paper/5e1d5c5d163da2e46150f9d4ff4e0340296865a1	urls	<ul style="list-style-type: none">http://arxiv.org/pdf/2103.11933v3http://arxiv.org/abs/2103.11933v3http://arxiv.org/pdf/2103.11933v3		
	id	id-1055901011345457162	id	id8354248203711072317		
	abstract	This study provides an efficient approach for using text data to calculate patent-to-patent (p2p) technological similarity, and presents a hybrid framework for leveraging the resulting p2p similarity for applications such as semantic search and automated patent classification. We create embeddings using Sentence-BERT (SBERT) based on patent claims. To further increase the patent embedding quality, we use transformer models based on SBERT and RoBERT, and apply the augmented approach for fine-tuning SBERT by in-domain supervised patent claims data. We leverage SBERTs efficiency in creating embedding distance measures to map p2p similarity in large sets of patent data. We deploy our framework for classification with a simple Nearest Neighbors (KNN) model that predicts Cooperative Patent Classification (CPC) of a patent based on the class assignment of the K patents with the highest p2p similarity. We thereby validate that the p2p similarity captures their technological features in terms of CPC overlap, and at the same demonstrate the usefulness of this approach for automatic patent classification based on text data. Furthermore, the presented classification framework is simple and the results easy to interpret and evaluate by end-users. In the out-of-sample model validation, we are able to perform a multi-label prediction of all assigned CPC classes on the subclass (663) level on 1,492,294 patents with an accuracy of 54% and F1 score > 66%, which suggests that our model outperforms the current state-of-the-art in text-based multi-label and multi-class patent classification. We furthermore discuss the applicability of the presented framework for semantic IP search, patent landscaping, and technology intelligence. We finally point towards a future research agenda for leveraging multi-source patent embeddings, their appropriateness across applications, as well as to improve and validate patent embeddings by creating domain-expert curated Semantic Textual Similarity (STS) benchmark datasets.	abstract	This study provides an efficient approach for using text data to calculate patent-to-patent (p2p) technological similarity, and presents a hybrid framework for leveraging the resulting p2p similarity for applications such as semantic search and automated patent classification. We create embeddings using Sentence-BERT (SBERT) based on patent claims. We leverage SBERTs efficiency in creating embedding distance measures to map p2p similarity in large sets of patent data. We deploy our framework for classification with a simple Nearest Neighbors (KNN) model that predicts Cooperative Patent Classification (CPC) of a patent based on the class assignment of the K patents with the highest p2p similarity. We thereby validate that the p2p similarity captures their technological features in terms of CPC overlap, and at the same demonstrate the usefulness of this approach for automatic patent classification based on text data. Furthermore, the presented classification framework is simple and the results easy to interpret and evaluate by end-users. In the out-of-sample model validation, we are able to perform a multi-label prediction of all assigned CPC classes on the subclass (663) level on 1,492,294 patents with an accuracy of 54% and F1 score > 66%, which suggests that our model outperforms the current state-of-the-art in text-based multi-label and multi-class patent classification. We furthermore discuss the applicability of the presented framework for semantic IP search, patent landscaping, and technology intelligence. We finally point towards a future research agenda for leveraging multi-source patent embeddings, their appropriateness across applications, as well as to improve and validate patent embeddings by creating domain-expert curated Semantic Textual Similarity (STS) benchmark datasets.		
	versions		versions			