	doc_1		doc_2		decision	id
cases	authors	• Leal, Rafael			NOT DUPLICATES	297
	title	Unsupervised zero-shot classification of Finnish documents using pre-trained language models	authors	• Yves Scherer		
	publication_date	2020-01-01 00:00:00	title	Unsupervised zero-shot classification of Finnish documents using pre-trained language models		
	source	SupportedSources.CORE	publication_date	None		
	journal		source	SupportedSources.SEMANTIC_SCHOLAR		
	volume		journal			
	doi	None	volume			
	urls	• https://core.ac.uk/download/362153774.pdf	doi			
	id	id-4553583199258622739	urls	• https://www.semanticscholar.org/paper/0ad7fa9d1db5a92666727530ecf3c80ab1bd8bcf		
	abstract	In modern Natural Language Processing, document categorisation tasks can achieve success rates of over 95% using fine-tuned neural network models. However, so-called "zero-shot" situations, where specific training data is not available, are researched much less frequently. The objective of this thesis is to investigate how pre-trained Finnish language models fare when classifying documents in a completely unsupervised way: by relying only on their general "knowledge of the world" obtained during training, without using any additional data. Two datasets are created expressly for this study, since labelled and openly available datasets in Finnish are very uncommon: one is built using around 5k news articles from Yle, the Finnish Broacasting Company, and the other, 100 pieces of Finnish legislation obtained from the Semantic Finlex data service. Several language representation models are built, based on the vector space model, by combining modular elements: different kinds of textual representations for documents and category labels, different algorithms that transform these representations into vectors (TF-IDF, Annif, fastText, LASER, FinBERT, S-BERT), different similarity measures and post-processing techniques (such as SVD and ensemble models). This approach allows for a variety of models to be tested. The combination of Annif for extracting keywords and fastText for producing word embeddings out of them achieves F1 scores of 0.64 on the Finlex dataset and 0.73-0.74 on the Yle datasets. Model ensembles are able to raise these figures by up to three percentage points. SVD can bring these numbers to 0.7 and 0.74-0.75 respectively, but these gains are not necessarily reproducible on unseen data. These results are distant from the ones obtained from state-of-the-art supervised models, but this is a method that is flexible, can be quickly deployed and, most importantly, do not depend on labelled data, which can be slow and expensive to make. A reliable way to set the input parameter for SVD would be an important next step for the work done in this thesis	id	id6397083763343697594			
			abstract	In modern Natural Language Processing, document categorisation tasks can achieve success rates of over 95% using fine-tuned neural network models. However, so-called â€œzero-shotâ€ situations, where specific training data is not available, are researched much less frequently. The objective of this thesis is to investigate how pre-trained Finnish language models fare when classifying documents in a completely unsupervised way: by relying only on their general â€œknowledge of the worldâ€ obtained during training, without using any additional data. Two datasets are created expressly for this study, since labelled and openly available datasets in Finnish are very uncommon: one is built using around 5k news articles from Yle, the Finnish Broacasting Company, and the other, 100 pieces of Finnish legislation obtained from the Semantic Finlex data service. Several language representation models are built, based on the vector space model, by combining modular elements: different kinds of textual representations for documents and category labels, different algorithms that transform these representations into vectors (TF-IDF, Annif, fastText, LASER, FinBERT, S-BERT), different similarity measures and post-processing techniques (such as SVD and ensemble models). This approach allows for a variety of models to be tested. The combination of Annif for extracting keywords and fastText for producing word embeddings out of them achieves F1 scores of 0.64 on the Finlex dataset and 0.73-0.74 on the Yle datasets. Model ensembles are able to raise these figures by up to three percentage points. SVD can bring these numbers to 0.7 and 0.74-0.75 respectively, but these gains are not necessarily reproducible on unseen data. These results are distant from the ones obtained from state-of-the-art supervised models, but this is a method that is flexible, can be quickly deployed and, most importantly, do not depend on labelled data, which can be slow and expensive to make. A reliable way to set the input parameter for SVD would be an important next step for the work done in this thesis. Acknowledgements This project was made possible by funding from the Finnish Ministry of Justice within the Anoppi project1 in partnership with the Semantic Computing Research Group (SeCo)2 at the University of Helsinki (HELDIG -Helsinki Centre for Digital Humanities) andAalto University. I would like to offer my special thanks to Yves Scherer from the University of Helsinki for his invaluable comments; Aki Hietanen and Tiina Husso, from the Ministry of Justice, and Eero HyvÃ¶nen, Jouni Tuominen, and the rest of the SeCo crew for their support. 1Automatic anonymisation and content description of documents containing personal data. https://oikeusmi nisterio.fi/en/project?tunnus=OM042:00/2018 2https://seco.cs.aalto.fi			
	versions		versions			