	doc_1	doc_2	decision	id
cases			DUPLICATES	26

authors	• Jacobs, Pieter Floris • Schomaker, Lambert • Wenniger, Gideon Maillette de Buy • Wiering, Marco
title	Active learning for reducing labeling effort in text classification tasks
publication_date	2021-09-10 00:00:00
source	SupportedSources.CORE
journal	
volume	
doi	None
urls	• https://core.ac.uk/download/489541336.pdf
id	id2124857306401389684
abstract	None
versions	

authors	• Jacobs, Pieter Floris • Schomaker, Lambert • Wenniger, Gideon Maillette de Buy • Wiering, Marco
title	Active learning for reducing labeling effort in text classification tasks
publication_date	2021-09-10 00:00:00
source	SupportedSources.CORE
journal	
volume	
doi	None
urls	• https://core.ac.uk/download/489541336.pdf
id	id2384629424618228397
abstract	Labeling data can be an expensive task as it is usually performed manually by domain experts. This is cumbersome for deep learning, as it is dependent on large labeled datasets. Active learning (AL) is a paradigm that aims to reduce labeling effort by only using the data which the used model deems most informative. Little research has been done on AL in a text classification setting and next to none has involved the more recent, state-of-the-art Natural Language Processing (NLP) models. Here, we present an empirical study that compares different uncertainty-based algorithms with BERT$_{base}$ as the used classifier. We evaluate the algorithms on two NLP classification datasets: Stanford Sentiment Treebank and KvK-Frontpages. Additionally, we explore heuristics that aim to solve presupposed problems of uncertainty-based AL; namely, that it is unscalable and that it is prone to selecting outliers. Furthermore, we explore the influence of the query-pool size on the performance of AL. Whereas it was found that the proposed heuristics for AL did not improve performance of AL; our results show that using uncertainty-based AL with BERT$_{base}$ outperforms random sampling of data. This difference in performance can decrease as the query-pool size gets larger.Comment: Accepted as a conference paper at the joint 33rd Benelux Conference on Artificial Intelligence and the 30th Belgian Dutch Conference on Machine Learning (BNAIC/BENELEARN 2021). This camera-ready version submitted to BNAIC/BENELEARN, adds several improvements including a more thorough discussion of related work plus an extended discussion section. 28 pages including references and appendice
versions	