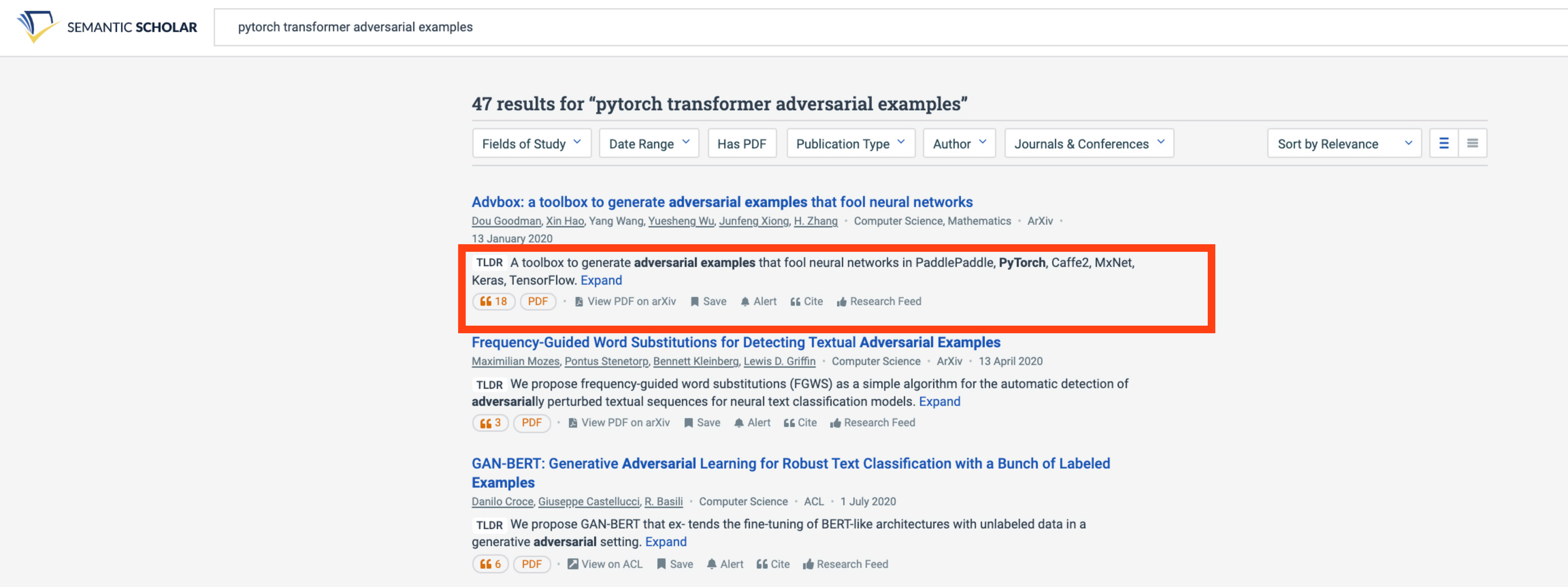

47 results for “pytorch transformer adversarial examples”

Advbox: a toolbox to generate adversarial examples that fool neural networks

Dou Goodman, Xin Hao, Yang Wang, Yuesheng Wu, Junfeng Xiong, H. Zhang • Computer Science, Mathematics • ArXiv • 13 January 2020

TLDR A toolbox to generate **adversarial examples** that fool neural networks in PaddlePaddle, **PyTorch**, Caffe2, MxNet, Keras, TensorFlow. Expand

Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples

Maximilian Mozes, Pontus Stenetorp, Bennett Kleinberg, Lewis D. Griffin • Computer Science • ArXiv • 13 April 2020

TLDR We propose frequency-guided word substitutions (FGWS) as a simple algorithm for the automatic detection of **adversarial**ly perturbed textual sequences for neural text classification models. Expand

GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples

Danilo Croce, Giuseppe Castellucci, R. Basili • Computer Science • ACL • 1 July 2020

TLDR We propose GAN-BERT that ex- tends the fine-tuning of BERT-like architectures with unlabeled data in a generative **adversarial** setting. Expand