	doc_1		doc_2		decision	id
cases	authors	• Di Jin • Zhijing Jin • Joey Tianyi Zhou • Peter Szolovits	authors	• Di Jin • Zhijing Jin • Joey Tianyi Zhou • Peter Szolovits	NOT DUPLICATES	433
	title	Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment	title	Is BERT Really Robust? Natural Language Attack on Text Classification and Entailment		
	publication_date	2019-07-27 00:00:00	publication_date	2019-07-27 00:00:00		
	source	SupportedSources.SEMANTIC_SCHOLAR	source	SupportedSources.SEMANTIC_SCHOLAR		
	journal		journal	ArXiv		
	volume		volume	abs/1907.11932		
	doi	10.1609/AAAI.V34I05.6311	doi			
	urls	• https://www.semanticscholar.org/paper/ae04f3d011511ad8ed7ffdf9fcfb7f11e6899ca2	urls	• https://www.semanticscholar.org/paper/a3347bbd82938788ec085772813c095de17a0b37		
	id	id-289308310322915074	id	id-274753565480961535		
	abstract	Machine learning algorithms are often vulnerable to adversarial examples that have imperceptible alterations from the original counterparts but can fool the state-of-the-art models. It is helpful to evaluate or even improve the robustness of these models by exposing the maliciously crafted adversarial examples. In this paper, we present TextFooler, a simple but strong baseline to generate adversarial text. By applying it to two fundamental natural language tasks, text classification and textual entailment, we successfully attacked three target models, including the powerful pre-trained BERT, and the widely used convolutional and recurrent neural networks. We demonstrate three advantages of this framework: (1) effectiveâ€”it outperforms previous attacks by success rate and perturbation rate, (2) utility-preservingâ€”it preserves semantic content, grammaticality, and correct types classified by humans, and (3) efficientâ€”it generates adversarial text with computational complexity linear to the text length.1	abstract	Machine learning algorithms are often vulnerable to adversarial examples that have imperceptible alterations from the original counterparts but can fool the state-of-the-art models. It is helpful to evaluate or even improve the robustness of these models by exposing the maliciously crafted adversarial examples. In this paper, we present the TextFooler, a general attack framework, to generate natural adversarial texts. By successfully applying it to two fundamental natural language tasks, text classification and textual entailment, against various target models, convolutional and recurrent neural networks as well as the most powerful pre-trained BERT, we demonstrate the advantages of this framework in three ways: (i) effectiveâ€”it outperforms state-ofthe-art attacks in terms of success rate and perturbation rate; (ii) utility-preservingâ€”it preserves semantic content and grammaticality, and remains correctly classified by humans; and (iii) efficientâ€”it generates adversarial text with computational complexity linear in the text length.		
	versions		versions			