		doc_1		doc_2	decision	id
cases	authors	David Herel and Hugo Cisneros and Tomas Mikolov	authors	Cisneros, Hugo Herel, David Mikolov, Tomas	DUPLICATES	117
	title	Preserving Semantics in Textual Adversarial Attacks	title	Preserving Semantics in Textual Adversarial Attacks		
	publication_date	2022-11-08 00:00:00	publication_date	2022-11-08 00:00:00		
	source	SupportedSources.INTERNET_ARCHIVE	source	SupportedSources.CORE		
	journal		journal			
	volume		volume			
	doi		doi	None		
	urls	https://web.archive.org/web/20221109152332/https://arxiv.org/pdf/2211.04205v1.pdf	urls	http://arxiv.org/abs/2211.04205		
	id	id8122764968160528802	id	id67701299540569289		
	abstract	Adversarial attacks in NLP challenge the way we look at language models. The goal of this kind of adversarial attack is to modify the input text to fool a classifier while maintaining the original meaning of the text. Although most existing adversarial attacks claim to fulfill the constraint of semantics preservation, careful scrutiny shows otherwise. We show that the problem lies in the text encoders used to determine the similarity of adversarial examples, specifically in the way they are trained. Unsupervised training methods make these encoders more susceptible to problems with antonym recognition. To overcome this, we introduce a simple, fully supervised sentence embedding technique called Semantics-Preserving-Encoder (SPE). The results show that our solution minimizes the variation in the meaning of the adversarial examples generated. It also significantly improves the overall quality of adversarial examples, as confirmed by human evaluators. Furthermore, it can be used as a component in any existing attack to speed up its execution while maintaining similar attack success.	abstract	Adversarial attacks in NLP challenge the way we look at language models. The goal of this kind of adversarial attack is to modify the input text to fool a classifier while maintaining the original meaning of the text. Although most existing adversarial attacks claim to fulfill the constraint of semantics preservation, careful scrutiny shows otherwise. We show that the problem lies in the text encoders used to determine the similarity of adversarial examples, specifically in the way they are trained. Unsupervised training methods make these encoders more susceptible to problems with antonym recognition. To overcome this, we introduce a simple, fully supervised sentence embedding technique called Semantics-Preserving-Encoder (SPE). The results show that our solution minimizes the variation in the meaning of the adversarial examples generated. It also significantly improves the overall quality of adversarial examples, as confirmed by human evaluators. Furthermore, it can be used as a component in any existing attack to speed up its execution while maintaining similar attack success.Comment: 8 pages, 4 figure		
	versions		versions			