Machine Learning and Knowledge Extraction (book)

Improving short text classification through global augmentation methods

Sefara, Tshephisho J

Council for Scientific and Industrial Research
Pretoria, 0001, South Africa
Email: TSefara@csir.co.za

Abstract

We study the effect of different approaches to text augmentation. To do this we use three datasets that include social media and formal text in the form of news articles. Our goal is to provide insights for practitioners and researchers on making choices for augmentation for classification use cases. We observe that Word2Vec-based augmentation is a viable option when one does not have access to a formal synonym model (like WordNet-based augmentation). The use of mixup further improves performance of all text based augmentations and reduces the effects of overfitting on a tested deep learning model. Round-trip translation with a translation service proves to be harder to use due to cost and as such is less accessible for both normal and low resource use-cases.