An overview of word-embedding methodologies to extinguish bias in deep-learning approaches to toxic comment classification

Hassan R. S. Andrabi

Abstract

Interactions occurring online are typically insensitive to the threat of social accountability for toxic behaviour, leading to frequent employment of abusive and anti-social tactics that go far beyond what might be considered acceptable in face-to-face settings. While applications of machine learning techniques to automatically detect and classify instances of toxic behaviour are well-studied, it has proven difficult to occlude biases in training datasets from flowing onwards to the classifier, and thereafter contributing to discriminatory classifications against sensitive classes such as race, religion, and gender. In this note, I seek to illuminate the capacity for word-embedding techniques to suppress undue learning of sensitive biases in training datasets. I consider this objective in the context of two popular deep-learning frameworks for toxic comment classification: long short-term memory (LSTM) networks, and convolutional neural networks (CNNs). I demonstrate that [...].

1 Introduction

The rise of the internet has transformed the primary setting of our social interactions to one which promotes extreme behaviours and dispenses fewer consequences. Ours is the age of cyberspace: now, more than ever before, individuals possess an unrestrained freedom to express their opinions for all to behold. This new setting does not reprimand us when we express opinions anarchically and without requisition. Instead, it promotes tendentious behaviour that is insular to empathetic considerations. It is against this contextual background that the majority of modern social interactions transpire — social interactions that exploit, by and large, the traceable anonymity of online systems and the resulting protection against social accountability. Indeed, moderation of online interactions is crucial to maintaining positive and healthy discussions. Naturally, this raises an important question: how do we efficiently and effectively evaluate the vast and inexorable flow of online interactions to limit proliferation of abusive and anti-social behaviour?

Given these analytical problems and the impending social importance of developing measures to promote healthy online interactions, the analysis in this note will empirically examine the mitigating capacity of machine-learning techniques to identify and filter out textual instances of toxic behaviour. To this end, I focus my analysis on the assessment of the joint classification capacity of: (1) a range popular of machine-learning techniques; and, (2) a suite of text-embedding representations. In particular, I consider the capacity of these model-embedding representations to minimise incidence of biased classifications: that is, the erroneous tendency for discriminatory classification against sensitive classes such as race, religion, and gender.