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

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras lectus arcu, gravida eget volutpat sit amet, finibus quis odio. Integer mattis turpis quis ante posuere, eu faucibus metus dictum. Morbi congue dui ultrices imperdiet ullamcorper. Praesent a nisl quis purus vestibulum vestibulum. Vestibulum eu lorem sed est tristique tincidunt. Aliquam a accumsan libero. Etiam erat mauris, egestas sit amet maximus ut, dapibus sed leo. Sed molestie, eros a finibus condimentum, massa dui sollicitudin nunc, ac maximus sem neque in nulla.

Duis eu condimentum dolor, eu viverra ligula. Morbi finibus dolor id risus pulvinar iaculis. Nulla facilisi. Nam et rutrum metus. Praesent fringilla placerat eros a cursus. Donec feugiat justo sed magna elementum venenatis. Proin at tempus lorem. Cras vestibulum augue volutpat elit tempor suscipit. Cras est risus, semper in lacinia in, condimentum non sem. Pellentesque mattis efficitur dapibus. Maecenas vulputate sit amet massa nec pretium. Quisque mattis, nisl in aliquet faucibus, lectus neque interdum massa, ut maximus diam purus at ante. Phasellus venenatis purus nulla. Nullam aliquet suscipit velit et dapibus. Nulla facilisi. Curabitur consectetur placerat erat at auctor.

Sed aliquam purus vitae urna dictum, eget semper eros pellentesque. Vestibulum consequat risus neque, consectetur vulputate erat iaculis id. Integer ultrices bibendum diam, id interdum dui. Vivamus feugiat velit id sodales vestibulum. Phasellus feugiat turpis dapibus

ante volutpat, a pulvinar quam volutpat. Nullam volutpat augue nisl, non sollicitudin augue sollicitudin in. Donec eget lectus dictum, ultricies ipsum sit amet, pellentesque mauris.