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.

2 Related Work

Prior research has already examined the capacity of various machine-learning and data-engineering techniques to classify toxic comments. Of general interest to the analysis is note is the literature on sentiment analysis, which combines natural-language-processing (NLP) techniques and opinion mining to emulate human-level comprehension of positive or negative sentiment expressed in textual statements [7, 8]. A relatively new subset of NLP literature considers applications of sentiment-analysis to the task of toxic behaviour. Research in this domain can be stratified with respect to the specific dimension of toxic behaviour of interest; besides general classification of toxic online comments [10, 22, 25], related literature considers more specific characterisiations of toxic behaviour, inluding hate speech [3, 19, 23, 27]; harrassessment [1, 4, 16]; abusive-language [6, 20, 26]; and cyber-bullying [2, 9, 14].

At the data pre-processing level, a number of studies consider potential for improved toxic-comment classification through pre-processing using sophisticated word-embedding techniques, such as TF-IDF [11, 15], GloVe [21], Word2Vec [17, 18], and FastText [5, 12, 13]. These techniques systematically estimate vector representations for words in a specified vocabulary, such that words arising from common contexts exhibit similar vector representations. Effective and well-studied word-embedding techniques (for a recent review, see Birunda and Devi, 2021 [24]) .

References

[1] S Abarna, JI Sheeba, S Jayasrilakshmi, and S Pradeep Devaneyan. Identification of cyber harassment and intention of target users on social media platforms. *Engineering applications of artificial intelligence*, 115:105283, 2022.

[2] Arnisha Akhter, Uzzal K Acharjee, and Md Masbaul A Polash. Cyber bullying detection and classification using multinomial naïve bayes and fuzzy logic. *Int. J. Math. Sci. Comput*, 5(4):1–12, 2019.

[3] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, and Idowu Ademola Osinuga. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, 38:100311, 2020.

[4] Priyam Basu, Tiasa Singha Roy, Soham Tiwari, and Saksham Mehta. Cyberpolice: Classification of cyber sexual harassment. In *EPIA Conference on Artificial Intelligence*, pages 701–714. Springer, 2021.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[6] Peter Bourgonje, Julian Moreno-Schneider, Ankit Srivastava, and Georg Rehm. Automatic classification of abusive language and personal attacks in various forms of online communication. In *International Conference of the German Society for Computational Linguistics and Language Technology*, pages 180–191. Springer, Cham, 2017.

[7] Erik Cambria and Bebo White. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57, 2014.

[8] KR1442 Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

[9] Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino. Unsupervised cyber bullying detection in social networks. In *2016 23rd International conference on pattern recognition (ICPR)*, pages 432–437. IEEE, 2016.

[10] Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, and Vassilis P Plagianakos. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th hellenic conference on artificial intelligence*, pages 1–6, 2018.

[11] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

[12] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[13] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[14] Tarek Kanan, Amal Aldaaja, and Bilal Hawashin. Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in arabic social media contents. *Journal of Internet Technology*, 21(5):1409–1421, 2020.

[15] Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 1957.

[16] Tolba Marwa, Ouadfel Salima, and Meshoul Souham. Deep learning for online harassment detection in tweets. In *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–5. IEEE, 2018.

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[19] Nanlir Sallau Mullah and Wan Mohd Nazmee Wan Zainon. Advances in machine learning algorithms for hate speech detection in social media: a review. *IEEE Access*, 2021.

[20] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[22] Julian Risch and Ralf Krestel. Toxic comment detection in online discussions. In *Deep learning-based approaches for sentiment analysis*, pages 85–109. Springer, 2020.

[23] Georgios Rizos, Konstantin Hemker, and Björn Schuller. Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 991–1000, 2019.

[24] S Selva Birunda and R Kanniga Devi. A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application*, pages 267–281, 2021.

[25] Betty Van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. *arXiv preprint arXiv:1809.07572*, 2018.

[26] Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300, 2020.

[27] Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the third workshop on abusive language online*, pages 11–18, 2019.