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

Ours is the age of cyberspace: now, more than ever before, individuals possess an unrestrained freedom to express their opinions for all to behold. Indeed, the rise of the internet has transformed the primary setting of our social interactions to one which promotes extreme behaviours and dispenses fewer consequences. 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 this paper 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 [8, 9]. A relatively new branch 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 [14, 25, 29], related literature also considers classification of specific dimensions of toxic behaviour, inluding hate speech [4, 19, 26, 31]; harrassessment [1, 6, 18]; abusive-language [7, 20, 30]; and cyber-bullying [2, 11, 16].

With respect to the selection of machine-learning frameworks employed for toxic comment classification, prior research is generally consistent in its advocation of certain machine learning techniques as better-suited for the task of toxic comment classifications. In particular, approaches towards classification of toxicity appear to prefer employment of convolutional neural networks (CNNs) [3], as these methods are known to excel at tasks involving elements of pattern-recognition. While CNNs are perhaps most well-known for applications in image-recognition [22], effective translation to sentiment-analysis tasks is unsurprising given the role of syntactical pattern recognition in language comprehension. For example, syntactical patterns such as word order, indiciative pharases, and idioms all modify meaning in a way that is algorithmic and theoretically 'learnable'. Indeed, understanding the implications of such patterns is essential to accurate comprehension of language.

At the data pre-processing level, a number of studies consider potential for improved toxic-comment classification through pre-processing using word-embedding techniques, such as TF-IDF [5, 10, 27], GloVe [5, 12, 24], Word2Vec [13], and FastText [24] (for a recent review of these techniques, see Birunda and Devi, 2021 [28]). These techniques systematically estimate vector representations for words in a specified vocabulary, such that words arising from common contexts exhibit similar vector representations. In general, most research in this field is hampered by challenges associated with skewed class distributions, leading to uneven training exposure to different classes of toxicity. To some extent, these challenges can be mitigated by intentional downsampling of training datasets to impose equal class distributions — although this comes at the cost of less overall data for model training.

3 Dataset

The analysis in this note uses the Jigsaw/Conversation AI Unintended Bias in Toxicity Classification competition dataset (available online: `https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/overview/description`). The dataset contains 155,000 annotated comments collected from an archive of the Civil Comments platform: a commenting plugin for online news sites. These comments were annoated by human raters usining binary toxicity labels, as well as a series of binary identity labels representing social identities mentioned in the comments. To obtain toxicity labels, each comment was presented to at least 10 human raters, who were prompted to rate comment toxicity according to predefined critera presented in Table 1. Notably, all comments included in this dataset were subject to a peer-review screening process imposed by Civil Comments. This manual peer-review system was designed to filter out obvious instances toxicity, and substantially limits diversity of vocabulary across the dataset. In particular, the dataset contains very few instances of profane language, and is unlikely to generalise effectively to contexts with less restrictive tenets of commenting etiquette. Figures 1 and 2 present visualisations of the most frequent words appearing in toxic and non-toxic comments respectively.

3.1. Pre-processing

Textual content of comments were cleaned and pre-processed using word-embedding techniques. In particular, all comments were normalised to lower-case, stripped of punctutation marks and non-alphabetic characters, and then tokenised into vector representations through word-embedding techniques. I apply pre-processing using three popular word-embedding represetations: term frequency - inverse document frequency (TF-IDF) [15, 17]; Global Vectors for Word Representation (GloVe) [21]; and Sentence-BERT [23]. Each of these techniques attempts to generate vector representations of textual content, such that sentences arising from similar contexts exhibit similar vector representations. Thereafter, the total dataset is partitioned into train and validation sets, with 140,000 and 15,000 instances allocated to each set respectively.

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] Darko Androcec. Machine learning methods for toxic comment classification: a systematic review. *Acta Universitatis Sapientiae, Informatica*, 12(2):205–216, 2020.

[4] 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.

[5] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760, 2017.

[6] 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.

[7] 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.

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

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

[10] Claudio Moisés Valiense de Andrade and Marcos André Gonçalves. Profiling hate speech spreaders on twitter: Exploiting textual analysis of tweets and combinations of multiple textual representations. In *CEUR Workshop Proc*, volume 2936, pages 2186–2192, 2021.

[11] 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.

[12] Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*, pages 105–114, 2019.

[13] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90, 2017.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[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] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[23] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[24] Alison Ribeiro and Nádia Silva. Inf-hateval at semeval-2019 task 5: Convolutional neural networks for hate speech detection against women and immigrants on twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 420–425, 2019.

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

[26] 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.

[27] Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. Hateminers: Detecting hate speech against women. *arXiv preprint arXiv:1812.06700*, 2018.

[28] 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.

[29] 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.

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

[31] 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.

List of Figures

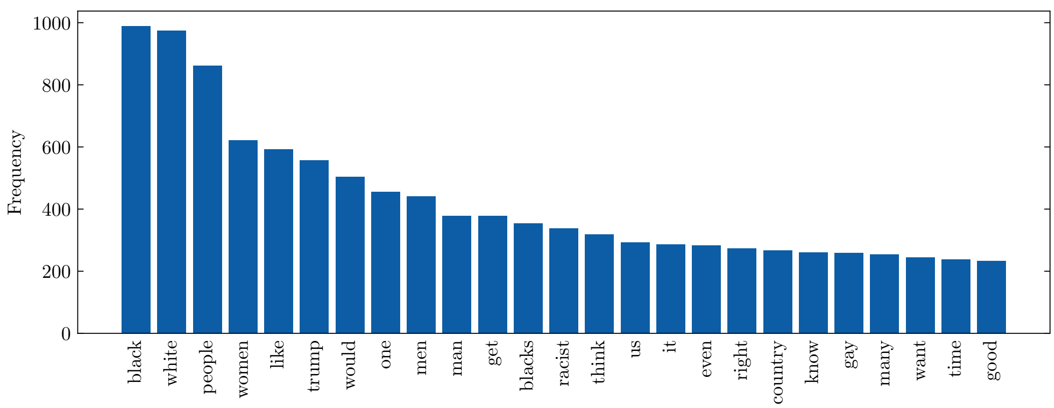

Figure 1: Words appearing most frequently in toxic comments. Frequencies represent the number of times an indicated word appears in a comment with a 'toxic' annotation. The top twenty-five most frequent words are labelled.

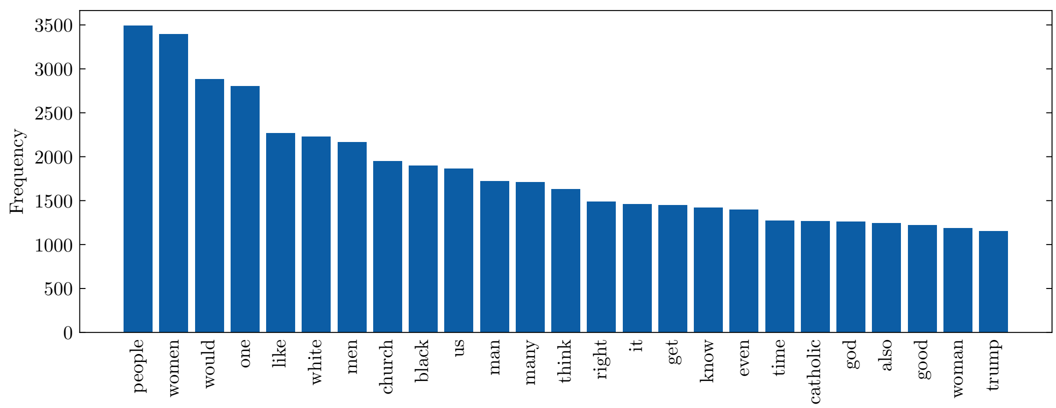

Figure 2: Words appearing most frequently in non-toxic comments. Frequencies represent the number of times an indicated word appears in a comment with a 'non-toxic' annotation. The top twenty-five most frequent words are labelled.

List of Tables

Table 1: Jigsaw/Coversation AI toxicity labelling criteria

Label	Criteria
Very Toxic	A very hateful, aggressive, or disrespectful comment that is very likely to make you leave a discussion or give up on sharing your perspective
Toxic	A rude, disrespectful, or unreasonable comment that is somewhat likely to make you leave a discussion or give up on sharing your perspective
Hard to say	No criteria given
Not toxic	No criteria given