A comparison of model-architecture and word-embedding methodologies to extinguish bias in 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. Automated machine learning techniques can detect and classify instances of toxic behaviour when they occur, however it has proven difficult to occlude biases in training datasets from flowing onwards to the classifier, resulting in discriminatory classifications against sensitive classes such as race, religion, and gender. In this paper, I seek to illuminate the capacity for combinations of model architecture and word-embedding techniques to suppress undue influence of such discriminatory classification. I consider this objective in the context of two popular perceptron-based approaches to toxic comment classification, and identify convolutional neural networks augmented with GloVe embeddings as the superior classifier.

1 Introduction

Ours is the age of cyberspace: now, more than ever before, individuals possess unrestrained freedom to anarchically express their opinions for all to behold. Freedom of speech is a desirably quality in systems of social interaction, though it cannot be denied that prevailing online culture champions polarising behaviours that are 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 afforded protection against social accountability. Indeed, the rise of the internet has transformed the primary setting of our social interactions: but can we be trusted to handle our newfound freedom in a responsible way? In the absence of naturally civilised interactions, moderation becomes crucial to maintaining positive and healthy discussions — yet, the volume of online communication is too vast to be effectively moderated in a manual fashion. This raises an important and imminent question: is it possible to automate supervision of the inexorable flow of online interactions in such a way that limit toxic 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 known to be accurate classifiers of toxic behaviour; and, (2) a suite of text-embedding representations for numerical encoding of textual data. In particular, I consider the capacity of these model-embedding combinations to minimise incidence of biased classifications: that is, the erroneous tendency for discriminatory classification against sensitive classes such as race, religion, and gender. The resulting experimental analysis intends to address the following research questions:

> **RQ1**: How does the choice of model and its architecture influence discriminatory biases in classification?
>
> **RQ2**: How does the choice of word-embedding techniques influence discriminatory biases in classification accuracy?
>
> **RQ3**: Which combination of model-architecture and word-embedding technique is the best-suited to toxic comment classification?

The remainder of the paper is structures as follows. Section 2 provides a contextual overview of relevant work in the field of toxic comment classification. Section 3 outlines the dataset and pre-processing methods employed in this analysis. Section 4 describes the general experimental framework, and the architecture of studied classification models. Section 5 presents and discusses the results. Finally, Section 6 concludes the paper.

2 Related Work

Prior research has 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 [9, 10]. A relatively new branch of NLP literature considers applications of sentiment-analysis to the task of toxic behaviour classification. Research in this domain can be grouped with respect to the specific dimension of toxic behaviour of interest: besides general classification of toxic online comments [15, 26, 30], related literature also considers classification of specific dimensions of toxic behaviour, including hate speech [5, 20, 27, 32]; harassment [1, 7, 19]; abusive-language [8, 21, 31]; and cyber-bullying [3, 12, 17].

Regarding selection of machine-learning frameworks for toxic comment classification, prior research is generally consistent in its advocation of certain techniques as better-suited for the task than others. In particular, prior research in toxicity classification prefers employment of convolutional neural networks (CNNs) [4], 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 [23], 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, indicative phrases, and idioms all

modify meaning in a way that is relatively 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 [6, 11, 28], GloVe [6, 13, 25], Word2Vec [14], and FastText [25] (see Birunda and Devi, 2021, for a review [29]). These techniques estimate vector representations for words in a specified vocabulary, such that words arising from common contexts exhibit similar vector representations.

3 Dataset

The analysis in this note employs 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`) [2]. The dataset contains 155,000 annotated comments collected from an archive of the Civil Comments platform: a commenting plugin for online news sites. Each comment within the dataset is annotated by human raters using 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 asked to rate comment toxicity according to predefined criteria 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 against profane communication. 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 a range of natural language processing techniques. In particular, all comments were cast to lower-case, stripped of punctuation marks and non-alphabetic characters, and then tokenised into vector representations through word-embedding techniques. I apply pre-processing using three popular word-embedding representations: term frequency - inverse document frequency (TF-IDF) [16, 18]; Global Vectors for Word Representation (GloVe) [22]; and Sentence-BERT [24]. Each of these techniques attempts to generate vector representations of textual content, such that sentences arising from similar contexts exhibit similar vector representations. Prior to GloVe based embedding, comments were padded or truncated to a token-length of 100, in order to ensure consistent dimensionality of inputs required for model estimation. Thereafter, the total dataset is partitioned into train and validation sets, with 140,000 and 15,000 instances allocated to each set respectively.

4 Experimental method

The following section outlines model architectures and estimation methodology employed to extract dependency between words and phrases in textual comments. To explicate further, consider the following example comment from the analysis dataset: "Muslims hate gays and want them dead". When viewed in isolation, individual words such as: "Muslims", and "gays" are not particularly indicative of toxic motivations — these words may appear in a variety of perfectly healthy discussions. Toxic motives online become discernible when words appear in particular combinations, such as "hate gays", or "want them dead". In such examples, it is clear that effective classification of intent requires an understanding of the encoding of syntactic patterns based on relative positions of critical words. Indeed, these are the patterns that word-embedding methodologies aim to capture.

Leveraging three independent word-embedding techniques introduced in Section 3, I estimate the classification performance of four popular machine-learning frameworks: (1) logistic regression; (2) shallow neural networks; (3) deep neural networks; and (4) convolutional neural networks. To evaluate the models, I employ a series of standard metrics used in classification tasks: accuracy, precision, recall, and F1 score. I assess model performance using the total dataset, and then separately across subsets of comments targeting particular identity subclasses. In total, my analysis estimates fourteen combinations of model structure and word-embedding, summarised in Table 2. All models were trained for a total length of ten epochs.

4.1. Logistic regression model

I implement a binary logistic regression model to predict toxicity. To account for proportionally low prevalence of 'toxic' labelled comments in the input dataset, penalties for toxicity class weights are set to be inversely proportional to the prevalence of classes in the input dataset. Subsequently, the estimation process imposes larger penalties for inaccurate classifications of 'toxic' labels, as compared to 'non-toxic' labels. L2 regularisation is applied to weights during the estimation process.

4.2. Shallow neural network (NN) model

Given substantial non-linear processing has already occurred in pre-trained sentence-BERT and GloVe word-embeddings, an extensive number of hidden layers may not be required to sufficiently capture remaining non-linearity in the input-output relationship. Accordingly, I implement a shallow neural network (NN) consisting of a pre-trained word-embedding layer; one fully-connected hidden layer with one-hundred and twenty-eight nodes activated by the Rectified Linear-unit (ReLU) activation function; and a single-node output layer activated by the sigmoid function.

4.3. Deep neural network (DNN) model

Deep neural networks are well-studied to exhibit exceptional performance in sentiment-analysis tasks. I implement a DNN consisting of a pre-trained word-embedding layer; two

fully-connected hidden layers with one-hundred and twenty-eight nodes each; and a single-node output layer with a sigmoid activation function. As with the NN, hidden nodes are configured with Rectified Linear-unit (ReLU) activation functions.

4.4. Convolutional neural network (CNN) model

I estimate a relatively simplistic convolutional neural network (CNN), consisting of a pre-trained word-embedding layer; a one-dimensional convolutional layer with a unit stride, and Glorot Normal kernel initialisation; a max-pooling layer with a unit-stride and pool-size equal to two; a single fully-connected hidden layer with one-hundred and twenty-eight nodes; and a single-node output layer activated by the sigmoid function. All nodes in hidden and convolutional layers are configured with Rectified Linear-unit (ReLU) activation functions.

5 Results

Model estimation and evaluation is conducted using the hold-out method, with a train-test split of N=140,000 and N=15,000, respectively. Each model was separately trained on the training dataset using TF-IDF, GloVe, and sentence-BERT word-embedding representations. Thereafter, classification performance was evaluated on a held-out validation dataset with consistent embeddings. I assess model performance by way of precision, recall, and F-score metrics between toxicity predictions passed by the model, and ground-truth annotations labelled by human raters using the criteria defined in Table 1.

Evaluation results for the entire validation dataset are presented in Table 3. Of the 15,000 comments in the validation dataset, 12,165 (81.1%) were truly non-toxic, and 2,835 (18.9%)were toxic. The CNN with GloVe embedding is the most efficient classifier across the range of tested model and word-embedding combinations. While most tested models were able to achieve accuracy in excess of a naive zero-rule prediction of the majority class, the magnitude of difference between achieved- and zero-rule performance was low across all experiments. Notably, the DNN with TF-IDF embeddings was unsuccessful in learning the TF-IDF embeddings, and performed worse than simple prediction of the majority class. Across the NN, DNN, and CNN model classes, those augmented with GloVe embeddings generate the highest F-scores, and consistently outperform alternative embedding methods. These results suggest that GloVe embeddings enable the most effective classifications of toxicity. However, a simplistic logistic regression model with sentence-BERT word-embeddings outperforms all but two other model-embedding combinations in the analysis in terms of F-score.

Putting aside general classification performance, stratifying model performance by identity class of comments reveals a number of interesting insights. In general, the majority of tested models are better classifiers of toxicity when references to religion, and gender are involved, and considerably worse when race and sexual orientation are involved. In congruence with previous research in this field, the CNN architectures with GloVe embeddings exhibit the best performance across all identity class subsets.

6 Conclusion

In this paper, I consider the application of various machine-learning architectures and model-embedding methodologies to the task of toxic-comment classification. Consistent with prior research, my analysis identifies CNNs supplemented with GloVe embeddings to be the superior architecture for this task. While all tested architectures suffer from an erroneous tendency for discriminatory classification on the basis of sensitive identity biases, the choice of CNN architectures and GloVe embeddings marginally limits the extent of these biases.

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] Jigsaw/Conversation AI. Jigsaw unintended bias in toxicity classification. `https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/overview/description`. Accessed: 2022-09-29.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

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

[25] 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.

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

[27] 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.

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

[29] 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.

[30] 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.

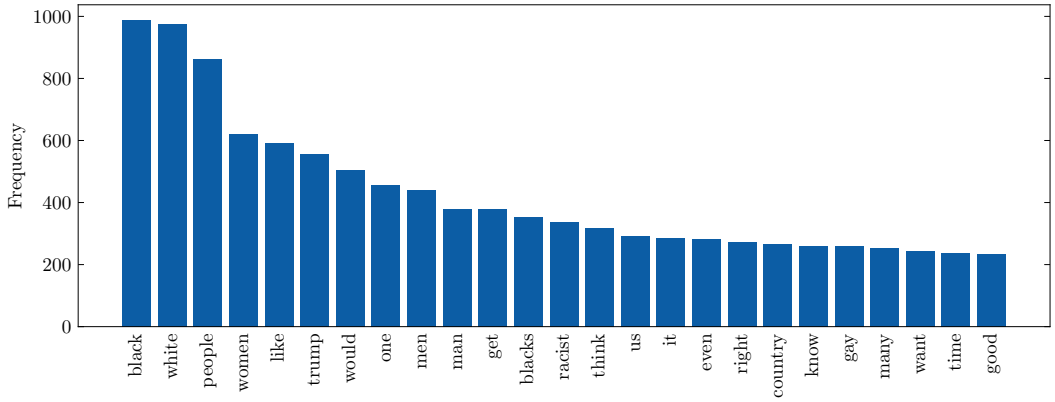
[31] Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300, 2020.

[32] 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

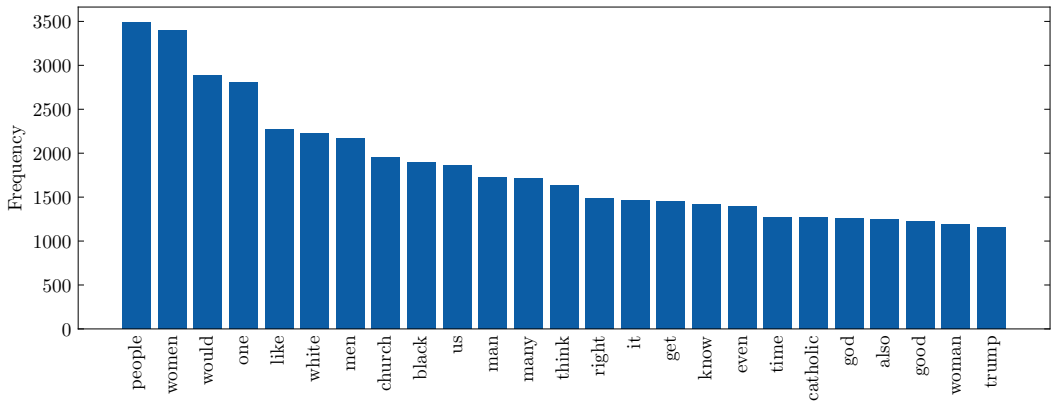
1 Words appearing most frequently in toxic comments 11
2 Words appearing most frequently in non-toxic comments 11

Figure 1: Words appearing most frequently in toxic comments



Notes: 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



Notes: 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

1 Jigsaw/Conversation AI toxicity labelling criteria 13
2 Model structure and word-embedding combinations 13
3 Toxicity classification performance 14
4 Toxicity classification performance for comments targeting religion 15
5 Toxicity classification performance for comments targeting sexual orientation 16
6 Toxicity classification performance for comments targeting race 17
7 Toxicity classification performance for comments targeting gender 18
8 Toxicity classification performance for comments targeting disability 19

Table 1: Jigsaw/Conversation 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

Table 2: Model structure and word-embedding combinations

Model	Word-embedding methodology
Logistic regression	TF-IDF embeddings
Logistic regression	sentence-BERT embeddings
Logistic regression	GloVe embeddings (minimum across 100 dimensions)
Logistic regression	GloVe embeddings (maximum across 100 dimensions)
Logistic regression	GloVe embeddings (average across 100 dimensions)
Shallow neural network (NN)	TF-IDF embeddings
Shallow neural network (NN)	sentence-BERT embeddings
Shallow neural network (NN)	GloVe embeddings
Deep neural network (DNN)	TF-IDF embeddings
Deep neural network (DNN)	sentence-BERT embeddings
Deep neural network (DNN)	GloVe embeddings
Convolutional neural network (CNN)	TF-IDF embeddings
Convolutional neural network (CNN)	sentence-BERT embeddings
Convolutional neural network (CNN)	GloVe embeddings

Table 3: Toxicity classification performance

Model	Precision	Accuracy	Recall	F-score
logit (TF-IDF)	0.6376	0.8276	0.2035	0.3086
logit (sentence-BERT)	0.6506	0.8335	0.2575	0.3690
logit (GloVe - min.)	0.5718	0.8143	0.0702	0.1250
logit (GloVe - max.)	0.5915	0.8165	0.0935	0.1614
logit (GloVe - avg.)	0.5833	0.8170	0.1111	0.1867
NN (TF-IDF)	0.6031	0.8215	0.1630	0.2566
NN (sentence-BERT)	0.6040	0.8254	0.2212	0.3238
NN (GloVe)	0.6438	0.8397	0.3404	0.4453
DNN (TF-IDF)	0.0000	0.8110	0.0000	0.0000
DNN (sentence-BERT)	0.6341	0.8295	0.2310	0.3387
DNN (GloVe)	0.7043	0.8380	0.2462	0.3649
CNN (TF-IDF)	0.6501	0.8235	0.1429	0.2342
CNN (sentence-BERT)	0.6235	0.8221	0.1478	0.2390
CNN (GloVe)	0.7011	0.8460	0.3228	0.4420

Notes: Model classification performance across entire validation dataset (N = 15,000). Logistic regression models are estimated with three distinct transformations of GloVe embeddings by coalescing vectors for all words in a given comment using the minimum, maximum, and average value across 100 dimensions.

Table 4: Toxicity classification performance for comments targeting religion

Model	Precision	Accuracy	Recall	F-score
logit (TF-IDF)	0.6607	0.8770	0.1729	0.2741
logit (sentence-BERT)	0.7143	0.8820	0.2025	0.3155
logit (GloVe - min.)	0.5909	0.8682	0.0607	0.1102
logit (GloVe - max.)	0.6173	0.8696	0.0779	0.1383
logit (GloVe - avg.)	0.6071	0.8707	0.1059	0.1804
NN (TF-IDF)	0.6333	0.8724	0.1184	0.1995
NN (sentence-BERT)	0.6686	0.8780	0.1822	0.2864
NN (GloVe)	0.6346	0.8832	0.3084	0.4151
DNN (TF-IDF)	0.0000	0.8657	0.0000	0.0000
DNN (sentence-BERT)	0.7239	0.8782	0.1511	0.2500
DNN (GloVe)	0.7194	0.8837	0.2196	0.3365
CNN (TF-IDF)	0.7108	0.8730	0.0919	0.1628
CNN (sentence-BERT)	0.6975	0.8755	0.1293	0.2181
CNN (GloVe)	0.7172	0.8878	0.2726	0.3950

Notes: Model classification performance across subset of columns targeting religion (N = 4,779). Comments were selected as those with a positive human-rated annotation in any of the following identity columns: Asian; Atheist; Buddhist; Christian; Hindu; Jewish; Muslim; or Other religion. Logistic regression models are estimated with three distinct transformations of GloVe embeddings by coalescing vectors for all words in a given comment using the minimum, maximum, and average value across 100 dimensions.

Table 5: Toxicity classification performance for comments targeting sexual orientation

Model	Precision	Accuracy	Recall	F-score
logit (TF-IDF)	0.6504	0.7756	0.2768	0.3883
logit (sentence-BERT)	0.6509	0.7711	0.2388	0.3494
logit (GloVe - min.)	0.8750	0.7480	0.0242	0.0471
logit (GloVe - max.)	0.7273	0.7560	0.0830	0.1491
logit (GloVe - avg.)	0.5814	0.7489	0.0865	0.1506
NN (TF-IDF)	0.6111	0.7711	0.3045	0.4065
NN (sentence-BERT)	0.6471	0.7694	0.2284	0.3376
NN (GloVe)	0.5750	0.7640	0.3183	0.4098
DNN (TF-IDF)	0.0000	0.7427	0.0000	0.0000
DNN (sentence-BERT)	0.6456	0.7631	0.1765	0.2772
DNN (GloVe)	0.6800	0.7667	0.1765	0.2802
CNN (TF-IDF)	0.7067	0.7703	0.1834	0.2912
CNN (sentence-BERT)	0.6333	0.7569	0.1315	0.2178
CNN (GloVe)	0.6726	0.7774	0.2630	0.3781

Notes: Model classification performance across subset of columns targeting sexual orientation (N = 1,123). Comments were selected as those with a positive human-rated annotation in any of the following identity columns: bisexual; heterosexual; homosexual gay or lesbian; and other sexual orientation. Logistic regression models are estimated with three distinct transformations of GloVe embeddings by coalescing vectors for all words in a given comment using the minimum, maximum, and average value across 100 dimensions.

Table 6: Toxicity classification performance for comments targeting race

Model	Precision	Accuracy	Recall	F-score
logit (TF-IDF)	0.6435	0.7229	0.2750	0.3853
logit (sentence-BERT)	0.6567	0.7335	0.3273	0.4369
logit (GloVe - min.)	0.5885	0.6948	0.1117	0.1878
logit (GloVe - max.)	0.6196	0.7005	0.1336	0.2198
logit (GloVe - avg.)	0.5898	0.6990	0.1539	0.2441
NN (TF-IDF)	0.6257	0.7155	0.2469	0.3541
NN (sentence-BERT)	0.6153	0.7177	0.2836	0.3882
NN (GloVe)	0.6409	0.7372	0.3820	0.4787
DNN (TF-IDF)	0.0000	0.6842	0.0000	0.0000
DNN (sentence-BERT)	0.6208	0.7254	0.3352	0.4353
DNN (GloVe)	0.6912	0.7355	0.2938	0.4123
CNN (TF-IDF)	0.6443	0.7128	0.2023	0.3080
CNN (sentence-BERT)	0.6262	0.7098	0.2016	0.3050
CNN (GloVe)	0.6924	0.7456	0.3500	0.4650

Notes: Model classification performance across subset of columns targeting race (N = 4,053). Comments were selected as those with a positive human-rated annotation in any of the following identity columns: Black; Latino; other race or ethnicity; or White. Logistic regression models are estimated with three distinct transformations of GloVe embeddings by coalescing vectors for all words in a given comment using the minimum, maximum, and average value across 100 dimensions.

Table 7: Toxicity classification performance for comments targeting gender

Model	Precision	Accuracy	Recall	F-score
logit (TF-IDF)	0.6439	0.8453	0.1441	0.2355
logit (sentence-BERT)	0.6286	0.8495	0.2195	0.3254
logit (GloVe - min.)	0.4667	0.8338	0.0356	0.0661
logit (GloVe - max.)	0.4882	0.8342	0.0525	0.0949
logit (GloVe - avg.)	0.5580	0.8383	0.1059	0.1781
NN (TF-IDF)	0.5911	0.8398	0.1017	0.1735
NN (sentence-BERT)	0.5710	0.8420	0.1771	0.2704
NN (GloVe)	0.6494	0.8597	0.3297	0.4373
DNN (TF-IDF)	0.0000	0.8347	0.0000	0.0000
DNN (sentence-BERT)	0.6147	0.8460	0.1839	0.2831
DNN (GloVe)	0.6862	0.8551	0.2280	0.3422
CNN (TF-IDF)	0.6557	0.8427	0.1017	0.1761
CNN (sentence-BERT)	0.5964	0.8407	0.1127	0.1896
CNN (GloVe)	0.6723	0.8602	0.3008	0.4157

Notes: Model classification performance across subset of columns targeting gender (N = 7,137). Comments were selected as those with a positive human-rated annotation in any of the following identity columns: female; male; other gender; or transgender. Logistic regression models are estimated with three distinct transformations of GloVe embeddings by coalescing vectors for all words in a given comment using the minimum, maximum, and average value across 100 dimensions.

Table 8: Toxicity classification performance for comments targeting disability

Model	Precision	Accuracy	Recall	F-score
logit (TF-IDF)	0.5610	0.8000	0.2255	0.3217
logit (sentence-BERT)	0.7500	0.8351	0.3235	0.4521
logit (GloVe - min.)	0.5000	0.7897	0.0294	0.0556
logit (GloVe - max.)	0.6429	0.7979	0.0882	0.1552
logit (GloVe - avg.)	0.4545	0.7876	0.0490	0.0885
NN (TF-IDF)	0.5263	0.7918	0.0980	0.1653
NN (sentence-BERT)	0.6458	0.8186	0.3039	0.4133
NN (GloVe)	0.6905	0.8227	0.2843	0.4028
DNN (TF-IDF)	0.0000	0.7897	0.0000	0.0000
DNN (sentence-BERT)	0.7027	0.8206	0.2549	0.3741
DNN (GloVe)	0.7692	0.8186	0.1961	0.3125
CNN (TF-IDF)	0.6667	0.8021	0.1176	0.2000
CNN (sentence-BERT)	0.8095	0.8165	0.1667	0.2764
CNN (GloVe)	0.7407	0.8433	0.3922	0.5128

Notes: Model classification performance across subset of columns targeting disability (N = 485). Comments were selected as those with a positive human-rated annotation in any of the following identity columns: intellectual or learning disability; other disability; physical disability; or psychiatric or mental illness. Logistic regression models are estimated with three distinct transformations of GloVe embeddings by coalescing vectors for all words in a given comment using the minimum, maximum, and average value across 100 dimensions.