A debiasing framework for deep learning applied to the morphological classification of galaxies

Esteban Medina-Rosales
Department of Computer Science
University of Concepción
Concepción, Chile
emedina2016@udec.cl

Guillermo Cabrera-Vives
Department of Computer Science, University of Concepción, Concepción, Chile
Millennium Institute of Astrophysics, Chile
guillecabrera@inf.udec.cl

Abstract

The morphologies of galaxies and their relation with physical features have been extensively studied in the past. Galaxy morphology labels are usually created by humans and are used to train machine learning models. Human labels have been shown to contain biases in terms of observational parameters such as the resolution of the labeled images. In this work, we demonstrate that deep learning models trained on biased galaxy data produce biased predictions. We also propose a method to train neural networks that takes into account this inherent labeling bias. We show that our deep de-biasing method is able to reduce the bias of the models even when trained using biased data.

1 Introduction

Astronomers have been studying galaxy properties and their evolution for over a century. Galaxy morphologies have been used to understand physical processes involved in their evolution, such as rotation velocity and gas content [8]. Large all-sky surveys such as the Sloan Digital Sky Survey (SDSS) [24] have enabled these kind of studies by giving access to millions of astronomical sources. In the future, we expect this number to grow to billions of galaxies when the Vera Rubin Observatory starts to map the sky through time and space [12]. Manually classifying galaxy images according to their morphology becomes an expensive task. Galaxy Zoo [16, 9] has been successful in creating hundreds of thousands of non-expert human labels using a crowdsourcing strategy. Human labels have been shown to be biased in terms of observable parameters: low resolution galaxies are biased towards smoother types because the fine structure of these galaxies can not be distinguished by human annotators [1, 3, 9, 4]. These labels have been used to train automatic classification models, particularly deep learning architectures which have shown to outperform previous approaches [6, 11, 7, 19, 22, 5].

Deep learning models have shown to learn human biases that are present in the training data [2, 17, 15, 18]. The question that naturally arises is: do deep learning models learn the human observational biases when classifying galaxies according to their morphologies? And if so, is it possible to train models that automatically remove those biases from the labels? In this work, we train a deep residual network [10] with biased and de-biased labels from Galaxy Zoo 2, and show that the observational bias present in human labeled galaxy morphologies datasets are learned by deep learning models that use them as training sets. We also propose a learning strategy to train these models so that they take

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

into account these biases, and show that we are able to diminish the bias of the models, even when training with biased labels.

2 Methodology

2.1 Labeling bias and de-biasing

Consider a supervised learning task over a human labeled dataset $\mathcal{D} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^N$ consisting of pairs of features and biased labels $(\mathbf{x}_i, \tilde{y}_i)$. We assume that each of these pairs are sampled independently from a data distribution $p_{\text{bias}}(\mathbf{x}, \tilde{y})$ defined over $\mathcal{X} \times \mathcal{Y}$. We will assume that for each biased label $\tilde{y}_i$ there exists a latent *ground truth* label $y_i \in \mathcal{Y}$ that we want to predict through a function $f_{\mathbf{w}} : \mathcal{X} \rightarrow \mathcal{Y}$ with parameters $\mathbf{w}$ to be fitted. Consider a biasing parameter α (e.g. the resolution of the labeled image). In this work we consider a maximum likelihood approach and leave a Bayesian analysis for future work. Consider a classification problem where $\mathcal{Y} = 1, \ldots, K$, the likelihood of the data given the model parameters and the biasing parameter is

$$
\begin{aligned}
p(\mathcal{D}|\mathbf{w}, \{\alpha_i\}_{i=1}^N) &= \prod_{i=1}^N p(\tilde{y}_i|\mathbf{x}_i, \mathbf{w}, \alpha_i), &(1)\\
&= \prod_{i=1}^N \sum_{y_i} p(\tilde{y}_i, y_i|\mathbf{x}_i, \mathbf{w}, \alpha_i), &(2)\\
&= \prod_{i=1}^N \sum_{y_i} p(\tilde{y}_i|y_i, \alpha_i) p(y_i|\mathbf{x}_i, \mathbf{w}), &(3)
\end{aligned}
$$

where the sum in Eq. 3 runs over the possible values of y_i, and we assumed that the biased labels $\tilde{y}_i$ only depends on the true labels y_i and the biasing parameters of each object α_i, while the true label is inferred from the features $\mathbf{x}_i$ and the parameters of the model $\mathbf{w}$.

Notice that $p(\tilde{y}|y, \alpha)$ models the biasing process by assuming the biased labels $\tilde{y}$ do not depend directly on the features $\mathbf{x}$. At the same time, $p(y|\mathbf{x}, \mathbf{w})$ models the dependance of y with $\mathbf{x}$, making $\tilde{y}$ and $\mathbf{x}$ conditionally independent given y. We model $p(y|\mathbf{x}, \mathbf{w})$ using a neural network with parameters $\mathbf{w}$. We train our model by minimizing the negative log-likelihood $-\log p(\mathcal{D}|\mathbf{w}, \{\alpha_i\}_{i=1}^N)$ (see Eq. 3).

2.2 Modeling galaxy morphology labeling bias

In this work, we consider a binary classification task where we classify galaxies between smooth or disk according to the first level of the Galaxy Zoo 2 classification tree [23]. Labels were defined by majority voting and galaxies that did not belong to smooth or disk classes were discarded. We use as biasing parameter α the resolution of the galaxy image measured as the angular Petrosian radius of the galaxy over the angular size of the point spread function (PSF).

We do not expect smooth galaxies to be confounded with disk galaxies, even for a poor resolution, hence we defined the probability $p(\tilde{y} = \text{disk}|y = \text{smooth}, \alpha) = 0$. On the other hand, we do expect low resolution disk galaxies to be confounded with smooth galaxies. This confusion will be more important for lower resolution images than for better resolved images. We follow [3], and model the biasing process as

$$p(\tilde{y} = \text{smooth}|y = \text{disk}, \alpha) = e^{-\alpha^2/(2\theta^2)}, \qquad (4)$$

where θ is a hyperparameter. Notice that $\lim_{\alpha \to 0} p(\tilde{y} = \text{smooth}|y = \text{disk}, \alpha) = 1$ (low resolution disk galaxies are always classified by humans as smooth) and $\lim_{\alpha \to \infty} p(\tilde{y} = \text{smooth}|y = \text{disk}, \alpha) = 0$ (high resolution disk galaxies are always clasified by humans as disks).

3 Results

In this work, we focus on Galaxy Zoo 2 (GZ2) data. We start by measuring the bias of the original crowdsourced labels and the de-biased labels produced by [9] (GZ2B and GZ2D hereafter). We train a ResNet50 model over these datasets and compare the bias on the predicted labels over the test set against the bias in the original labels of the same test set. We then use GZ2B to train a ResNet50 using our approach as described in Sec. 2. We also run the de-biasing procedure described in [3], which is based on a logistic regression model that takes as input intrinsic parameters of the galaxies (CV14 hereafter).

3.1 Data

All our experiments use GZ2 data [23]. We created hard labels by using majority voting and discarded "star or artifacts" objects. Aiming at using a balanced dataset, we randomly sampled 121,984 galaxies from which 62,225 corresponded to "smooth", and 59,759 corresponded to "features or disk" according to their original crowdsourced labels. In order to train the neural network models, we divided the dataset into a training set of 97,600 galaxies, a validation set of 12,192 galaxies, and a test set of 12,192 galaxies. We use the JPEG images labeled by the annotators as input to our models. All galaxy parameters needed in our experiments were obtained from the SDSS database.

3.2 Deep learning architecture and training

We used a ResNet50 [10] as our deep learning model with an extra dense layer (1024 neurons and ReLU activation) after the convolutional and pooling layers. We also trained Deep Galaxy V2 [13] and ResNet101 models, but report our results with ResNet50 given that it obtained the best performance when training over GZ2B and GZ2D (minimizing the cross entropy i.e. no de-biasing). We performed optimization by using ADAM [14] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. Table 1 shows the accuracy, precision, recall, and f1-score for these experiments.

Table 1: Biases for different datasets and experiments. Classification metrics are included for ResNet50 models trained directly over the GZ2B and GZ2D datasets.

Dataset / Method	Bias CV14	Bias CV18	accuracy	precision	recall	f1-score
a) GZ2B [23]	0.4831	0.3701	-	-	-	-
b) GZ2D [9]	**0.2679**	0.3086	-	-	-	-
c) ResNet50 over GZ2B	0.4893	0.3795	0.921	0.921	0.921	0.921
d) ResNet50 over GZ2D	0.2969	0.3246	0.896	0.866	0.879	0.872
e) CV14 [3]	0.3823	0.2990	-	-	-	-
f) DDB (ours)	0.3382	**0.2873**	-	-	-	-

3.3 Biases and de-biasing

In order to assess the level of bias in each set of labels, we use the bias quantities described in [3] and [4] (CV14 and CV18, respectively). CV14 assumes that the fractions of objects of each class in an unbiased dataset should not be significantly different for labels coming from images with different resolutions. Hence, they calculate the deviation of the fraction of objects for bins in α as compared to their intrinsic fraction. CV18 extends this idea by considering that the fractions of objects of each class should not significantly change for labels with different resolutions *within bins of intrinsic parameters*. By doing this, the intrinsic fractions of objects per class is assumed to vary in terms of the intrinsic parameters. Notice that the biases calculated in CV14 are not comparable to the ones in CV18, but they are useful when comparing the amount of bias in different labeled datasets by using each of them separately and considering their different assumptions.

All debiasing methods were applied directly to GZ2B, and used as biasing parameter the galaxy resolution as measured by $\alpha = (r/\mathrm{PSF})$, where r is the angular Petrosian radius of the galaxy and PSF is the angular standard deviation of a Gaussian model PSF. For the sake of comparison, we start by training a logistic regression model using the de-biasing procedure described in CV14 using the Sérsic index, the ellipticity, and the half-light radius [21] as features for the classifier. This procedure not only estimates the parameters of the logistic regression, but it also estimates de parameter θ of the bias distribution of Eq. 4. This value for GZB2 was estimated as $\theta = 9.18$.

As explained in Sec. 2 our proposed deep de-biasing method (DDB hereafter) consists of minimizing the negative logarithm of the likelihood described in Eq. 3. We use a ResNet50 and Adam optimization

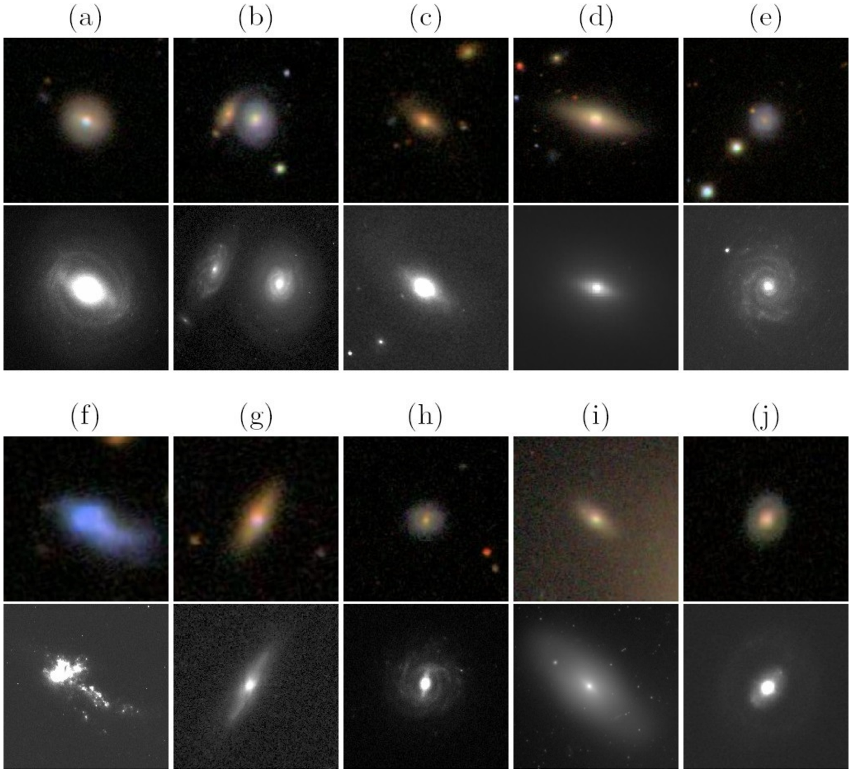

Figure 1: Low and high resolution images of galaxies that changed their labels from "smooth" to "disk" when using our method. The first rows show the images as labeled by the annotators of Galaxy Zoo 2. The second row shows higher resolution images from the HST.

with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. For the bias distribution we use the same distribution as in CV14 (Eq. 4) with a bias parameter of $\theta = 9.18$.

Table 1 shows the results of our experiments. We start by noticing that, as expected, the labels in GZ2D have a lower bias than the ones in GZ2B. The biases on the ResNet50 trained directly over GZ2B (Tab. 1c) and GZ2D (Tab. 1d) show a similar behaviour indicating that when training the ResNet50 model over biased data, the predicted labels will be as biased as the training set. In other words, training on biased data produces biased models. On the other hand, both CV14, and DDB are able to diminish the amount of bias of the original dataset GZ2B. Using DDB we are able to remove even more bias than CV14. Furthermore, CV14 relies on the calculation of the Sérsic profiles [20], while DDB is able to automatically learn features that help infer labels with less bias than the original dataset.

In order to visually assess the performance of our approach, we searched the Mikulski Archive for Space Telescopes[1] (MAST) for high resolution Hubble Space Telescope (HST) images of galaxies in the test set that changed their labels when using DDB. We found 10 HST images of galaxies that changed from a "smooth" biased label to a "disk" de-biased label and none that changed from "disk" to "smooth". Figure 1 shows the SDSS and HST images of such galaxies. Figures 1a, 1b, 1e, 1h, and 1j show evidence of spiral arms. Figure 1f shows irregular features. Figures 1c, 1d, 1g, 1i show lenticular features, although it is not completely clear that these are effectively disk galaxies. A further astrophysical analysis is required to corroborate their type, e.g. by using colors and magnitudes. These results show that our debiasing neural network is able to correctly label images even when the labels used for training are biased.

[1] https://mast.stsci.edu/portal/Mashup/Clients/Mast/Portal.html

4 Conclusions

We have studied the effect of training deep learning models for predicting galaxy morphologies using data that is biased in terms of observable parameters such as the resolution of the images. We have shown that when directly training these models using the biased data, training converges to a biased model. We proposed a method for training deep learning models using biased data. We showed that our method is able to converge to a model that diminishes the bias of the predicted labels.

Broader Impact

This work proposes a deep learning approach for estimating the ground truth labels of morphologically classified galaxies using a biased labeled dataset. We believe that having a better estimate of these labels will help astronomers to have a more detailed understanding of the formation and evolution of galaxies. Although this paper is focused on biases related to the observational process of astronomical image labeling, our approach may be applied to other kind of biases by modeling the distribution of biased labels $p(\tilde{y}|y, \alpha)$ according to the specific problem to be addressed. In practice, α may be any biasing parameter present in real world datasets. One of the drawbacks of our approach is that the biased label does not depend directly on the features used to infer the unbiased labels, which needs to be addressed in future work.

Acknowledgments and Disclosure of Funding

We gratefully acknowledge support from ANID through the FONDECYT Initiation grant Nº11191130 and from the Data Science Unit at University of Concepción. G.C.V. acknowledges support from the ANID –Millennium Science Initiative Program–ICN12_009 awarded to the Millennium Institute of Astrophysics (MAS). We are greateful for valuable insights from Pierluigi Cerulo and Monserrat Martínez.

References

[1] S. P. Bamford, R. C. Nichol, I. K. Baldry, K. Land, C. J. Lintott, K. Schawinski, A. Slosar, A. S. Szalay, D. Thomas, M. Torki, D. Andreescu, E. M. Edmondson, C. J. Miller, P. Murray, M. J. Raddick, and J. Vandenberg. Galaxy Zoo: the dependence of morphology and colour on environment. *Monthly Notices of the Royal Astronomical Society*, 393:1324–1352, March 2009.

[2] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018.

[3] Guillermo F. Cabrera, Chris J. Miller, and Jeff Schneider. Systematic labeling bias: De-biasing where everyone is wrong. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, page in press. IEEE, 2014.

[4] Guillermo Cabrera-Vives, Christopher J. Miller, and Jeff Schneider. Systematic labeling bias in galaxy morphologies. *The Astronomical Journal*, 156(6):284, nov 2018.

[5] Ting-Yun Cheng, Christopher J Conselice, Alfonso Aragón-Salamanca, Nan Li, Asa F L Bluck, Will G Hartley, James Annis, David Brooks, Peter Doel, Juan García-Bellido, David J James, Kyler Kuehn, Nikolay Kuropatkin, Mathew Smith, Flavia Sobreira, and Gregory Tarle. Optimizing automatic morphological classification of galaxies with machine learning and deep learning using Dark Energy Survey imaging. *Monthly Notices of the Royal Astronomical Society*, 493(3):4209–4228, 02 2020.

[6] Sander Dieleman, Kyle W Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459, 2015.

[7] H Domínguez Sánchez, M Huertas-Company, M Bernardi, D Tuccillo, and J L Fischer. Improving galaxy morphologies for SDSS with Deep Learning. *Monthly Notices of the Royal Astronomical Society*, 476(3):3661–3676, 02 2018.

[8] A. Dressler. Galaxy morphology in rich clusters - Implications for the formation and evolution of galaxies. *The Astrophysical Journal*, 236:351–365, March 1980.

[9] Ross E Hart, Steven P Bamford, Kyle W Willett, Karen L Masters, Carolin Cardamone, Chris J Lintott, Robert J Mackay, Robert C Nichol, Christopher K Rosslowe, Brooke D Simmons, et al. Galaxy zoo: comparing the demographics of spiral arm number and a new method for correcting redshift bias. *Monthly Notices of the Royal Astronomical Society*, 461(4):3663–3682, 2016.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] M. Huertas-Company, R. Gravet, G. Cabrera-Vives, P. G. Pérez-González, J. S. Kartaltepe, G. Barro, M. Bernardi, S. Mei, F. Shankar, P. Dimauro, E. F. Bell, D. Kocevski, D. C. Koo, S. M. Faber, and D. H. Mcintosh. A Catalog of Visual-like Morphologies in the 5 CANDELS Fields Using Deep Learning. *Astrophysical Journal Supplement Series*, 221:8, November 2015.

[12] Željko Ivezić et al. LSST: From science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, mar 2019.

[13] Nour Eldeen Khalifa, Mohamed Hamed Taha, Aboul Ella Hassanien, and Ibrahim Selim. Deep galaxy v2: Robust deep convolutional neural networks for galaxy morphology classifications. In *2018 International Conference on Computing Sciences and Engineering (ICCSE)*, pages 1–6, 2018.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[16] Chris J. Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179, 2008.

[17] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3150–3158. PMLR, 10–15 Jul 2018.

[18] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), July 2021.

[19] M. Pérez-Carrasco, G. Cabrera-Vives, M. Martinez-Marin, P. Cerulo, R. Demarco, P. Protopapas, J. Godoy, and M. Huertas-Company. Multiband galaxy morphologies for CLASH: A convolutional neural network transferred from CANDELS. *Publications of the Astronomical Society of the Pacific*, 131(1004):108002, aug 2019.

[20] J. L. Sersic. *Atlas de galaxias australes*. Córdoba: Obs. Astronómico, 1968.

[21] L. Simard, J. T. Mendel, D. R. Patton, S. L. Ellison, and A. W. McConnachie. A Catalog of Bulge+disk Decompositions and Updated Photometry for 1.12 Million Galaxies in the Sloan Digital Sky Survey. *Astrophysical Journal Supplement Series*, 196:11, September 2011.

[22] Mike Walmsley, Lewis Smith, Chris Lintott, Yarin Gal, Steven Bamford, Hugh Dickinson, Lucy Fortson, Sandor Kruk, Karen Masters, Claudia Scarlata, Brooke Simmons, Rebecca Smethurst, and Darryl Wright. Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning. *Monthly Notices of the Royal Astronomical Society*, 491(2):1554–1574, 10 2019.

[23] Kyle W Willett, Chris J Lintott, Steven P Bamford, Karen L Masters, Brooke D Simmons, Kevin RV Casteels, Edward M Edmondson, Lucy F Fortson, Sugata Kaviraj, William C Keel, et al. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, page stt1458, 2013.

[24] Donald G. York and the SDSS Collaboration. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579–1587, sep 2000.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 2.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2.
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The code is not included. Reference to the data and instructions to reproduce the main results are described in Section 3.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 3.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] The same dataset is used for all experiments with no randomness involved.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] Can be reviewed by following the reference.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] Too extensive to be included. This information can be reviewed by following the reference.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] All participants are volunteers.