REMOVING RACIAL BIAS IN TED TALK RATINGS BY AWARENESS OF VERBAL AND GESTURE QUALITY

Rupam Acharyya [*]
Department of Mathematics
University at Buffalo
rupamach@buffalo.edu

Ankani Chattoraj [*]
Department of Brain & Cognitive Science
University of Rochester
achattor@ur.rochester.edu

Shouman Das
Department of Mathematics
University of Rochester
anumoshsad@gmail.com

Md. Iftekhar Tanveer
Spotify Research
go2chayan@gmail.com

Ehsan Hoque
Department of Computer Science
University of Rochester
mehoque@gmail.com

ABSTRACT

The role of verbal and non-verbal cues towards great public speaking has been a topic of exploration for years. We identify a commonality across present theories, the element of "variety or heterogeneity" in modes of communication (e.g. resorting to stories, scientific facts, emotional connections, facial expressions etc.) which is essential for effectively communicating information. Based on this observation, we formalize a novel HEterogeneity Metric, HEM, that quantifies the quality of a talk both in the verbal and non-verbal domain (transcript and facial gestures). Using TED talks as an input repository of public speeches, we show that there is a meaningful relationship between HEM and the ratings of TED talks given to speakers by viewers. Further, we discover that HEM successfully captures the prevalent bias in ratings w.r.t race. We thus incorporate the HEM metric into the loss function of a neural network and show improvement of fairness in rating prediction. Our work ties together a novel metric for public speeches in both verbal and non-verbal domain to design a fair rating prediction system.

1 INTRODUCTION

A great talk depends on how efficiently the inner thoughts of the speakers are expressed to an audience Costello (1930); Newcomb (1917). It requires encompassing ethos, pathos and logos Rapp (2010) as well as including variety, such as, humor, questions, quotations, analogies etc Sandmann (2013); Fisher (1989); Garlick (1993); Bitzer (1999); Ting-Toomey & Dorjee (2018); DeCaro et al. (2012). The nonverbal components of a talk also play a key role in determining its appeal Lucas & Stob (2004). Effective use of both transcript and gesture and a deliberate alteration of message through verbal and nonverbal cues give shape to the main message of a speech Abbott (2008); Tanveer et al. (2018). One common thread that ties the important aspects of a good public speech in both verbal and non-verbal domain is "variety or heterogeneity" which we formalize by defining a novel, "HEtero-geneity Metric" (HEM). We conduct our investigation on a diverse set of talks and ratings as found in the TED talk website. First, we show that desirable positive ratings (undesirable negative ratings) of talks grow (shrink) with the increase in HEM. This emphasizes that HEM indeed quantifies the quality of a talk based on heterogeneity/variety. However, we also observe that too much variability (higher values of HEM) can be overwhelming and distracting for the audience

[*] equal contribution

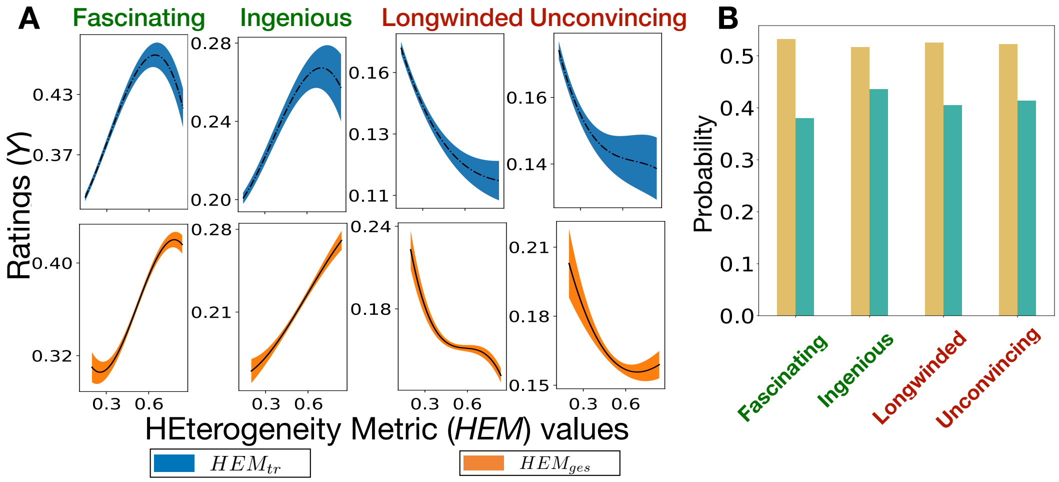

Figure 1: **A**: Positive ratings (green titles) show a concave/increasing trend for both HEM_{tr} (blue) and HEM_{ges} (orange). However, too much heterogeneity can lead to confusion causing positive ratings to saturate or even decrease. The trend is opposite for negative rating labels (red titles), showing a convex or decreasing relationship. **B**: Bias in rating probability in real data.

causing a decrease (increase) in positive (negative) ratings. Interestingly, we also find that HEM captures discrepancies in rating of TED talks w.r.t race. Motivated by this observation, we introduce *fairness by quality* by incorporating HEM into the loss function of a neural network and thus building a fair rating predictor for public speeches w.r.t. *race*.

2 Data

Data Acquisition: We collected the public speaking data from TEDtalk website [1] (1980 talks covering $>$ 400 categories and millions of viewers). Viewers rate the talks based on multiple labels of which we consider 2 positive/desirable rating labels: *fascinating, ingenious* and 2 negative/undesirable rating labels: *long-winded, unconvincing*. We collected data on the sensitive attribute S (*race*) using Amazon mechanical turk following Acharyya et al. (2020). We use information about the data attributes X: *transcript (Tr), visual gestures (V), view counts (C), sensitive attribute S: race (R), and rating label (Y)* of each talk to design our prediction model (see Table 1).

Data Preprocessing: We use transcript (verbal) and visual gestures (non-verbal) for construction of the heterogeneity based metric HEM. We obtain the doc2vec Le & Mikolov (2014) representation of each transcript ($Tr \in \mathbb{R}^d$) by using the `gensim` library Řehůřek & Sojka (2010) ($d = 200$) to compute HEM_{tr}. We normalize view counts of the TED talks ($C \in \mathbb{Z}$) using the min-max technique to obtain $C \in \mathbb{R}$ and make it comparable across attributes. We normalize rating labels Y for each talk by dividing with its total number of rating counts. Binary rating labels are computed by thresholding w.r.t its median across talks, $Y^{bin} \in \{0, 1\}$ (see Table 2 in appendix for details).

Choice of Rating Labels: We assume that heterogeneity adds "x-factor" to a speech and determines how engaged and appealing the speech is to an audience (used as the term "quality" associated with HEM). Based on that, we chose 2 positive and 2 negative rating labels which cannot be otherwise trivially attributed to common identifiable causes.

3 Heterogeneity Metric (HEM)

A good talk has credibility, emotional connection, logic, stories, scientific facts, quotations, humor etc which we call *characteristics*. Each characteristic involves different words specific to their nature and various facial expressions relevant to their associated emotion.

[1] Videos on `www.ted.com`

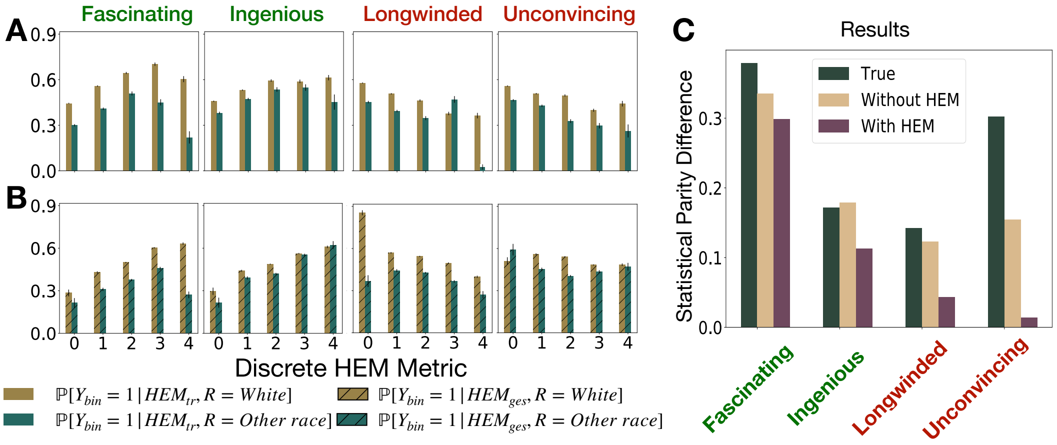

Figure 2: **A**: Significant difference between probability of obtaining a rating for White speakers than for speakers of Other race w.r.t HEM_{tr}^{dis} (trend agrees with true data in Figure 1 B). **B**: Same as **A** but for HEM_{ges}^{dis}. The non overlapping 95% CI shown in **A** and **B** indicate significance of our claim ($p < 0.05$). **C**: Fairness in prediction improves as quantified by lower values of SPD measure for race when HEM is used in the loss function of neural network.

Verbal (HEM_{tr}): For each talk we obtained K topics using Latent Dirichlet Allocation (LDA) Blei et al. (2003). Each of these topics is a probability distribution over the set of words in the transcript. We define the summary representation of each topic as the weighted combination (weights are the respective probability) of the Glove embeddings Pennington et al. (2014) for the highest probable words in that topic, i.e., the summary representation $\boldsymbol{T_i} (\in \mathbb{R}^{300})$ of i^{th} topic is defined by, $\boldsymbol{T_i} = \sum_{j=1}^{n} weight(w_j) \cdot glove(w_j)$, where n is the number of highest probable words chosen to represent the topic and $weight(w_j)$ is the weight of the j^{th} word in the i^{th} topic. We then define the representation matrix of all topics as $\boldsymbol{T} (\in \mathbb{R}^{K \times 300})$ where the i^{th} row is $\boldsymbol{T_i}$. We obtain the topic similarity matrix, $\boldsymbol{U} = \boldsymbol{T}\boldsymbol{T}^{\top} (\in \mathbb{R}^{K \times K})$ which captures the similarity between topics. $HEM_{tr} = \prod_{i=1}^{k} \lambda_i$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K \geq 0$ are the eigenvalues of $\boldsymbol{U}$. This product defines the volume spanned by the most diverse k topics in the topic vector space (see Theorem 5.2 of Kulesza et al. (2012)). Note that, the more diverse the topic vectors, larger will be the magnitude of the product of top k eigenvalues. Hence more variety and heterogeneity in a talk w.r.t the transcript implies greater value of HEM_{tr}. In this work, we choose $n = 10, K = 10, k = 5$.

Non-verbal (HEM_{ges}): We model the facial gestures similar to Agarwal et al. (2019) which encodes personalized speaking pattern. OpenFace Baltrušaitis et al. (2016) toolkit is used to extract 17 facial action units related intensity scores from the video at 30 frames/sec. These intensity scores represent various facial muscle movements like inner brow raiser, outer eyebrow raiser, eyebrow lowerer, upper lid raiser, cheek raiser, jaw drop, eye blink etc. Each talk video is divided into 10 second segments with 5 seconds sliding window. For each of these segments, we calculate the pairwise correlations between 17 facial action units (total of $\binom{17}{2} = 136$ correlation coefficients) to quantify the similarity among these intensity scores over time. We define the heterogeneity metric for a speaker's gesture pattern (HEM_{ges}) as the maximum distance between two points in 136-dimensional space. Intuitively, this maximum distance gives us the measure of highest variability present in facial muscle movements. We denote the number of 10 seconds segments in a TED talk video by s, $p_i (\in \mathbb{R}^{136})$ represents correlation coefficients for i^{th} segment, d_{ij} denotes the eucledian distance between p_i and p_j. The HEM_{ges} is defined as, $HEM_{ges} = \max_{i,j \in \{1,\cdots,s\}} d_{ij}$.

Relationship between HEM and Ratings: We validate that HEM is an useful metric by investigating its relation with ratings. We normalized the HEM values using min-max technique such that they lie in $[0, 1]$. These values are then divided into 5 equal sized bins and their corresponding mean rating is reported in the X-axis of Figure 1A. The Y-axis denotes the corresponding rating value. We find that positive ratings (here, *ingenious, fascinating*) show a concave/increasing trend and negative rating labels (here, *longwinded, unconvincing*) show convex/decreasing trend. This agrees with the psychological intuition that motivated definition of HEM and confirms that it indeed meaningfully influence ratings for both the verbal and non verbal regime.

Bias in Rating Captured by HEM: TED talks are rated by spontaneous viewers who come from different background and age groups. As a result we find "societal bias" in ratings w.r.t race. We find discrepancy between $\mathbb{P}[Y_{bin} = 1|\text{ R } = \text{ White }]$ and $\mathbb{P}[Y_{bin} = 1|\text{ R } = \text{Other race }]$ for a rating of interest R (Figure 1B). Bias in data can also be counter intuitive (see Figure 1 B). It is often expected that white speakers are less likely to be rated with negative labels when compared to speakers of other race. However, the data shows opposite trend, highlighting bias here is quantified as any type of discrepancy in rating probability irrespective of assumptions.
Since HEM captures the "heterogeneity based quality" of the talk, the ratings should be similar for similar quality values, irrespective of race. $HEM \in [0, 1]$ is discretized into 5 values, $HEM^{dis} \in 0, 1, 2, 3, 4$ by binning $[0, 1]$ into 5 equal sized bins. We then compute $\mathbb{P}[Y_{bin} = 1|\text{ R } = \text{ White }, HEM^{dis}]$ and $\mathbb{P}[Y_{bin} = 1|\text{ G } = \text{ Other race }, HEM^{dis}]$ for a fixed value of HEM^{dis} as shown in Figure 2 A and B. We find that speakers have significant discrepancy across all positive and negative rating labels for both transcript and gesture w.r.t race. The nature of the bias w.r.t. HEM^{dis} matches existing bias in data. The mismatch in bias pattern for higher HEM^{dis}_{tr} and lower HEM^{dis}_{ges} can be explained by negligible probability mass of HEM^{dis}_{tr} and HEM^{dis}_{ges} for values of 4 and 0 respectively (see Figure 4 in Appendix).

4 FAIR MODEL BASED ON MULTI-MODAL HEM

We design a neural network model whose loss function enforces the reduction of rating differences for similar HEM values. Note that the neural network has information about race in the input only and the rating discrepancy is minimized solely w.r.t. HEM.

Problem Formulation and Methods: Verbal feature of a TED talk is a 200 dimensional vector obtained from the doc2vec representation (Mikolov et al., 2013). We get the non-verbal feature by computing top k (in our case $k = 2$) eigenvalues for the correlation matrix of a segment (See Section 3 for detail about segment) and concatenate these eigenvalues across all the segments. We make this into a fixed length vector by padding 0's at the end. We also have the speaker's race information. Input $X^{(i)}(i \in \{1, \cdots, N\})$ is now created by concatenating the verbal feature, non-verbal feature, the sensitive attribute race and view counts. We use the binarized ratings $Y^{(i)}_{bin} = \left(y^{(i)}_1, \cdots, y^{(i)}_6\right)$ as the label. We also have HEM for each TEDtalk as defined in Section 3. Let us denote the overall heterogeneity by $h^{(i)} = (HEM^{(i)}_{tr}, HEM^{(i)}_{ges})$.
Our goal is to learn a prediction function f_θ such that its predicted output is not only close to the ground truth but also has similar prediction for inputs with similar HEM values. Formally we need to minimize the following loss function,

$$L(\theta) = Pred(\theta) + \lambda \cdot HEM(\theta) \tag{1}$$

where, $Pred(\theta) = \frac{1}{N}\sum_{i=1}^{N} BCE\left(f_\theta(X^{(i)}), Y^{(i)}\right)$ and $HEM(\theta) = \frac{1}{\binom{N}{2}}\sum_{i,j}\left\|\hat{Y}^{(i)} - \hat{Y}^{(j)}\right\|_2^2 \cdot \mathbb{I}\left(|h^{(i)} - h^{(j)}| < \epsilon\right)$ where BCE denotes the binary cross entropy loss and $\hat{Y}^{(i)} = f_\theta(X^{(i)})$. Here $Pred(\theta)$ represents the prediction loss w.r.t the ground truth and $HEM(\theta)$ controls the discrepancy of the prediction between inputs with similar HEM metric. Here ϵ and λ are the hyperparameters, where ϵ is the tolerance of HEM difference for which discrepancy in rating is penalized and λ controls the strength of HEM loss.

Results: We used a feed forward neural network with one hidden layer (400 hidden nodes) for classification. We use model accuracy to measure the performance of the model and SPD Biddle (2006) to quantify the fairness, where SPD measures the difference of rating predictions between two groups. Formally, SPD $= |\mathbb{P}(y_i = 1|R \in R_1) - \mathbb{P}(y_i = 1|R \in R_2)|$, where, $R_1 =$ White and R_2 = Other race. Smaller values of SPD indicates better fairness. Note that the SPD decreased after incorporating HEM into the loss function which indicates fair rating prediction. The average accuracy of all 6 rating labels is 67% when we train our model without any HEM loss ($\epsilon = \lambda = 0$). After adding the HEM loss to our model ($\epsilon = 0.02$ and $\lambda = 4.0$) we observe 64% accuracy and significant improvement in fairness w.r.t. race (Figure 2C). Note that, the model's accuracy is decreased by 3%, when trained with HEM in the loss function. The goal of this work is not to beat state of the art in prediction accuracy but to achieve improved fairness in prediction for comparable accuracy levels.

5 DISCUSSION

This work combines fairness in the rating of public speeches with the quality of a speech based on heterogeneity in verbal and non-verbal aspects. Though our focus was mostly on verbal and non-verbal aspects of the speeches, our work can be generalized to other modes. We only investigate two cases, white and other race, to establish the existence of bias in ratings and effectiveness of our paradigm. All other combinations of race can be explored in future work (also, gender). The simplicity of our encoding systems for defining HEM also makes it applicable to other similar datasets. Our setup can be used in hiring or recruiting systems. A company often looks for employees with diverse experience and abilities. Our novel HEM metric can be used to rate job talks given by interviewers. Similarly, this setup can be used for podcast ratings and other such systems.

REFERENCES

H Porter Abbott. *The Cambridge introduction to narrative*. Cambridge University Press, 2008.

Rupam Acharyya, Shouman Das, Ankani Chattoraj, and Md Iftekhar Tanveer. Fairyted: A fair rating predictor for ted talk data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 338–345, 2020.

Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 38–45, 2019.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10. IEEE, 2016.

Dan Biddle. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd., 2006.

Lloyd F Bitzer. The rhetorical situation. *Human Communication: Core Readings*, pp. 303–16, 1999.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Charles S Costello. A psychological approach to public speaking. 1930.

Peter DeCaro, Tyrone Adams, and Bonnie Jefferis. audience analysis. *Preuzeto*, 9:2018, 2012.

Walter R Fisher. Human communication as narration: Toward a philosophy of reason, value, and action. 1989.

Rick Garlick. Verbal descriptions, communicative encounters and impressions. *Communication Quarterly*, 41(4):394–404, 1993.

Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196, 2014.

Stephen Lucas and Paul Stob. *The art of public speaking*. McGraw-Hill New York, 2004.

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

Charles M Newcomb. The educational value of expression. *Quarterly Journal of Speech*, 3(1): 69–79, 1917.

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

Christof Rapp. Aristotle's rhetoric. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2010 edition, 2010.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

Warren Sandmann. introductions & conclusions. *Public Speaking: The Virtual Text*, pp. 1–12, 2013.

M Iftekhar Tanveer, Samiha Samrose, Raiyan Abdul Baten, and M Ehsan Hoque. Awe the audience: How the narrative trajectories affect audience perception in public speaking. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 24. ACM, 2018.

Stella Ting-Toomey and Tenzin Dorjee. *Communicating across cultures*. Guilford Publications, 2018.

Table 1: Breakdown of TED talk Dataset w.r.t. race

Property	**Sub Property**	**Quantity**
Total Number of talks		1980
Race of Speaker	White	1573
	African American	147
	Asian	173
	Other Race	87

Table 2: Pre-processed TED Dataset Attributes

Sensitive attributes S	R: race
Data attributes X	Tr: transcript V: visual gestures and C: view count
Label Y	ratings (3 positive and 3 negative) and Y: normalized ratings, Y^{bin}: binarized ratings

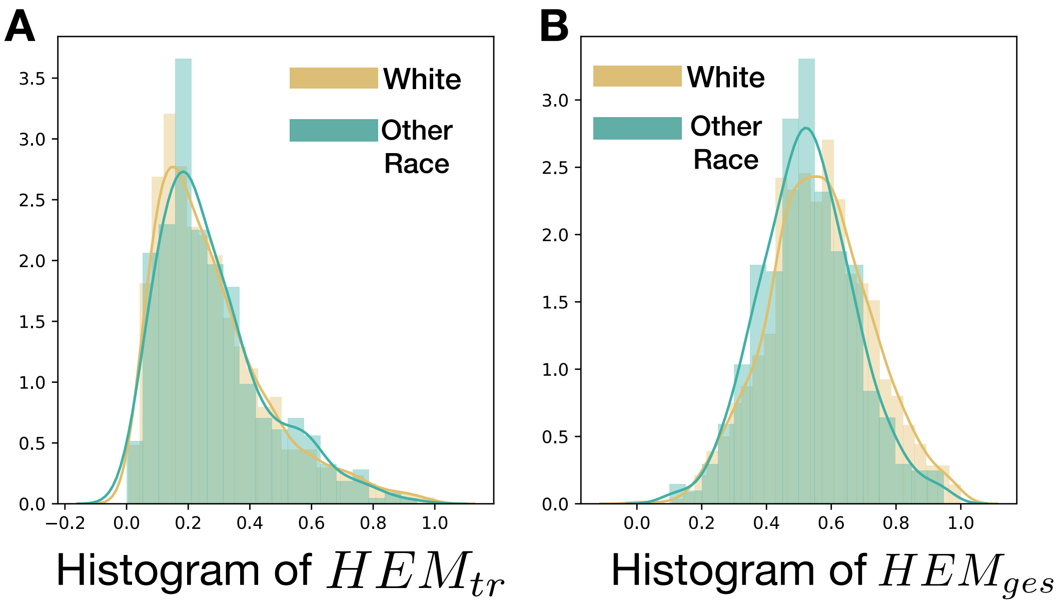

Figure 3: HEM_{tr} **and** HEM_{ges} **are not biased by definition**
A-B: Shows histogram of HEM_{tr} and HEM_{ges} for white speakers (yellow) and speakers of other race (green) for both HEM_{tr} and HEM_{ges} respectively. We find no significant difference in the two histograms which indicates that the bias observed in Figure 1 B are not an artifact of a biased metric.

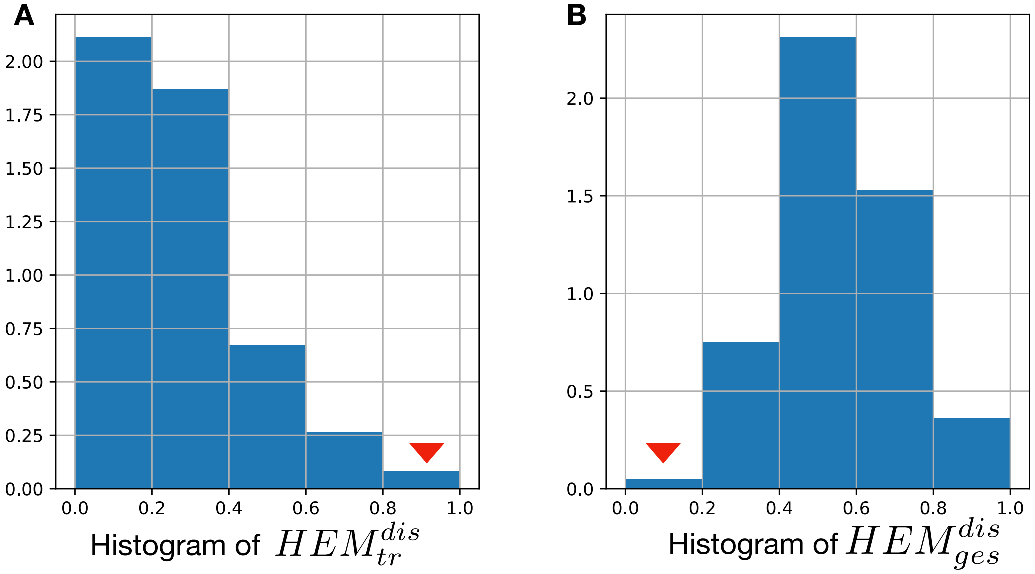

Figure 4: **Distribution of discretized HEM_{tr}^{dis} and HEM_{ges}^{dis}**
A: HEM_{tr} is discretized by assigning values between $[0.0, 0.2)$ to category 0, $[0.2, 0.4)$ to category 1 and so on $[0.8, 1.0]$ to 4. We find negligible instances for $HEM_{tr}^{dis} = 4$, (i.e, $HEM_{tr} >=$ 0.8 and $<= 1.0$) as shown in red. **B**: Similarly, $HEM_{ges}^{dis} = 0$, (i.e, $HEM_{ges} >= 0.0$ and < 0.2) is almost non-existent, pointed by red arrow.