Towards Automatic Bias Detection in Knowledge Graphs

Daphna Keidar[1]*, Mian Zhong[1] *, Ce Zhang[1], Yash Raj Shrestha[1], Bibek Paudel[2]
[1] ETH Zürich, Switzerland; [2] Stanford University, U.S.A.
{dkeidar, mzhong, yshrestha}@ethz.ch
ce.zhang@inf.ethz.ch, yshrestha@ethz.ch, bibekp@cs.stanford.edu

Abstract

With the recent surge in social applications relying on knowledge graphs, the need for techniques to ensure fairness in KG based methods is becoming increasingly evident. Previous works have demonstrated that KGs are prone to various social biases, and have proposed multiple methods for debiasing them. However, in such studies, the focus has been on debiasing techniques, while the relations to be debiased are specified manually by the user. As manual specification is itself susceptible to human cognitive bias, there is a need for a system capable of quantifying and exposing biases, that can support more informed decisions on what to debias. To address this gap in the literature, we describe a framework for identifying biases present in knowledge graph embeddings, based on numerical bias metrics. We illustrate the framework with three different bias measures on the task of profession prediction, and it can be flexibly extended to further bias definitions and applications. The relations flagged as biased can then be handed to decision makers for judgement upon subsequent debiasing.

1 Introduction

Knowledge graphs (KGs) update and represent world knowledge in a structured and scalable format. They are commonly embedded into lower dimensional representations, namely knowledge graph embeddings (KGEs), which have successfully been applied in diverse applications such as personalized recommendations (Liu et al., 2019), question answering (Huang et al., 2019), and enhancement of language modeling (Zhang et al., 2019; Peters et al., 2019; Baumgartner et al., 2018). Following the proliferation of social applications relying on KGEs, the issue of fairness in KG based methods is a growing concern.

Recent works show that KGEs are inclined to manifest bias, and propose methods for debiasing them (Fisher et al., 2020a; Arduini et al., 2020; Bose and Hamilton, 2019). However, these works implicitly assume that the relations to be debiased are chosen by the practitioner without quantification (e.g. based on social preconception), which may result in a sub-optimal decision. Reaching an informed decision on what to debias thus poses a challenge, and there is currently no empirical, data-driven method for identifying biased relations in KGs. In lack of such a system, some potential biases may go unnoticed while others are exaggerated.

In this paper we aim to fill this gap, and present a framework for numerically identifying biases in KGEs. Our goal is to facilitate decision making by providing a table of bias scores on KG relations, as well as to encourage exploratory research in comprehending the nature of KGE biases. Practically, we describe and implement the framework using three bias measures that we derive from bias definitions from the domain of machine learning fairness.[1] The relations and their corresponding bias scores can then guide practitioners when deciding which relations to debias.

We experiment and evaluate our framework's feasibility by implementing it for three bias definitions and applying it to the two benchmark datasets FB15K-237 (Bordes et al., 2013) and Wikidata5m (Wang et al., 2019).

2 Why Automatic Bias Detection?

2.1 Aptly Deciding what to Debias

Before debiasing a KGE, the user must choose which relations to debias. However, there is currently no method for uncovering which relations are prone to bias in a KG, and in previous works

*The first two authors contributed equally.

[1] The code will be released at: https://github.com/mianzg/kgbiasdetec

this selection was done mostly manually. Such manual selection of the relations to debias by the user can in itself be biased; for instance, the bias in a KG may potentially be rooted in non obvious relations such as ZIP codes (Krieger et al., 2002) or a person's given name which can go unnoticed. It is therefore imperative to measure biases across a broad set of possibly sensitive relations, in an extensive and empirical manner.

2.2 Identifying the Sources of Bias

Biases in KGEs can arise from multiple sources, including the data collection process for the KG, the chosen ontology, or the embedding method (Janowicz et al., 2018). To help better understand the sources of bias, our bias measurement framework can be used to measure bias in different embeddings of the same KG. By doing so, we can examine which biases are apparent or amplified in certain embedding approaches, analyse whether the embedding method affects bias, and infer which biases are inherent to the KG itself. The output of our framework opens the door to further studies comparing biases across embedding methods and KGs, and serves as a step towards uncovering the bias sources.

2.3 Comparing Bias Types

Moreover, the machine learning literature includes a range of bias and fairness definitions (Verma and Rubin, 2018). By extracting bias scores that are based on diverse bias definitions, we can empirically compare and analyse the relationships and correlations among them.

3 Our Framework for KGE Bias Identification

In this section, we begin with an overview of three specific bias measures we employ, which are defined over the relations in a KG. We then provide an overall description of our pipeline.

3.1 Preliminaries: Bias Measures

There is a multitude of definitions for fairness and bias in the machine learning literature (Verma and Rubin, 2018), and impossibility theorems have shown that they cannot all be simultaneously achieved (Saravanakumar, 2020; Kleinberg et al., 2016). In our model, we implemented three different definitions of fairness, that we formally describe below. Our framework can be easily extended to additional fairness definitions.

The first two measures are Predictive and Demographic Parity (Mitchell et al., 2021; William Dieterich, 2016), both common fairness metrics, which rely on a classification task. They measure the bias of sensitive relations via classification on a target relation. In our experiments, the classifier is trained to predict the target relation "profession" in order to measure bias in other relations.

Predictive Parity focuses on the classifier's precision, whereas Demographic Parity is useful when the underlying ground truth data is biased. These metrics are not specific to KGs and we describe below how their definition is extended to the KG setting. The third measure, Translational Likelihood Bias (TLB), is specifically tailored towards KGEs (Fisher et al., 2020b). It leverages the score function used in KGE training to update entity embeddings and compute bias.

Formally, a knowledge graph (KG) is a set of facts represented by triples of the form $(h, r, t) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, where $\mathcal{E}$ denotes the set of entities and $\mathcal{R}$ denotes the set of relations. Each triple (h, r, t) has a head entity h, a relation r, and a tail entity t, represented by embedding vectors. As we are concerned with fairness, we focus on the sub-graph containing solely human entities $\mathcal{H}$, and their associated relations as $R_{\mathcal{H}}$. There may exist a set of sensitive relations $S \subset R_{\mathcal{H}}$ related to humans, towards which we want to detect any biases. For Predictive and Demographic Parity, we also assume a classification task and a classifier which takes entity embeddings as input, together with a relation and predicts the corresponding tails. Lastly, we refer s as a possibly sensitive relation in the following definitions, and take it as binary for simplicity.

Demographic Parity A classifier satisfies demographic parity with respect to a sensitive relation s, if the classifier's predictions, denoted $\hat{y}$, are independent of s. Namely, demographic parity holds if $\mathbb{P}[\hat{y} = a|s = 1] = \mathbb{P}[\hat{y} = a|s = 0]$ for all possible predictions a. We can then measure the demographic parity distance (DPD) as

$$\text{DPD}(s, a) = |P[\hat{y} = a|s = 1] - P[\hat{y} = a|s = 0]| \quad (1)$$

We finally compute

$$\text{DPD}(s) = \sum_{a} \text{DPD}(s, a) \quad (2)$$

In the case of profession prediction, $\hat{y}$ stands for the predicted profession, and a stands for a possible profession. Intuitively, DPD measures how

	Demographic Parity Distance ↑				Predictive Parity Distance ↑				Translational Likelihood ↑			
	TransE	ComplEx	DistMult	RotatE	TransE	ComplEx	DistMult	RotatE	TransE	ComplEx	DistMult	RotatE
gender	0.127	0.177	0.191	0.182	0.037	0.008	0.012	0.01	**1.566**	0.047	**0.208**	0.010
languages	0.190	0.201	0.227	0.211	0.0	0.0	0.0	0.0	0.824	**0.293**	0.077	**0.012**
nationality	**0.280**	**0.361**	**0.289**	**0.277**	**0.267**	**0.296**	**0.215**	**0.368**	0.724	0.289	0.076	0.007

Table 1: Aggregated bias scores of three implemented measures when predicting profession in FB15K-237. The arrow indicates the direction of larger bias. We measure the bias scores with respect to gender, languages, and nationality across four different embedding methods. This table suggests an investigation into debiasing nationality, as it has the highest bias scores in all embeddings.

much the sensitive relation affects classification, and therefore depends on the data itself. For instance, let the sensitive relation s represent "is-Christian". In this setting, if all photographers in the KG were Christian, an accurate profession classifier would have a high DPD with respect to being Christian, and a random one would have low DPD. Namely, DPD rewards a classifier that is agnostic to the sensitive relation, regardless of its performance.

Predictive Parity A classifier satisfies this definition with respect to a sensitive relation s if its precision is independent of s. Given a predicted label a, the predictive parity distance (PPD) is then defined as

$$\begin{aligned} \text{PPD}(s,a) = |P[y=a|\hat{y}=a, s=1] \\ - P[y=a|\hat{y}=a, s=0]| \end{aligned} \tag{3}$$

We then compute

$$\text{PPD}(s) = \sum_a \text{PPD}(s,a) \tag{4}$$

Translational Likelihood KGEs are typically trained using a score function ϕ that is unique to the embedding method, and captures closeness between entities and relations in the embedding space. According to Fisher et al. (2020b), we can directly use such score functions to measure bias. Given a triple (h, r, t), we obtain a translated triple (h', r, t) by performing a one-step gradient descent to update the head entity embedding on a direction of sensitive relation s. For instance, if s is gender, we can translate the head entity h in the direction of the entity "female" to obtain a new h'. The translational likelihood bias is then calculated as the difference in scores between the original and translated triples on a target relation r_i

$$\text{TL}(s, t_{r_i}) = \phi(h', r_i, t_{r_i}) - \phi(h, r_i, t_{r_i}) \tag{5}$$

Following the example, a positive bias value indicates that making h more "female" results in a higher score for the tail occupation t_{r_i}, and a negative bias value indicates that the new entity h' scores worse on t_{r_i}. A score closer to zero suggests a relatively fairer relation, i.e. less bias.

Compared with DPD and PPD which are based on a downstream classification task, TLB does not require an external task to compute bias as it is calculated directly from the KGE.

3.2 Score Aggregation

While the scores in section 3.1 can be used to calculate bias for a binary sensitive relation, in practice, a sensitive relation s may have multiple tail values t. In general, given a possibly sensitive relation s, we are interested in a score function $\text{bias}_{\text{agg}}(s) : \mathcal{R}_{\mathcal{H}} \rightarrow \mathbb{R}$ that will summarise how much bias there is towards $s \in \mathcal{R}_{\mathcal{H}}$.

To generalize bias_{agg} to the non-binary case, consider a sensitive relation s and let $t_1, ...t_n$ be all the possible tails s can have. We aggregate the score over t_i;

$$\text{bias}_{\text{agg}}(s) = \frac{1}{n}\sum_{i=1}^{n} \text{bias}(s = t_i) \tag{6}$$

where $\text{bias}(s = t_i)$ is the bias with respect to having tail t_i with the relation s, and is calculated according to the measure of choice.

3.3 Bias Detection Framework

The workflow of our framework allows the user to specify a pre-trained KGE and a set of bias measurements. In the case of DPD and PPD, the user should also specify a classification task, namely the target relation to be classified. For clarity, throughout this paper we will consider profession to be the classification target. The output of our model is a table containing bias scores for each specified relation and bias measurement. This provides users with a ranking of relations according to their bias scores, which can empirically inform decisions on which relations to debias. After observing the bias scores for each relation, the practitioner can choose which ones to debias by according to their domain knowledge. Furthermore, our tool offers a multi-bias perspective for comparing bias across different embeddings, and helps select the appropriate embeddings according to downstream applications.

4 Experiments

We applied our framework with the bias measurements described in section 3.1 to evaluate bias on two benchmark datasets; FB15k-237 (Bordes et al., 2013) and Wikidata5m (Wang et al., 2019). Each of the datasets was trained on four KGE methods respectively; TransE, CompleEx, DistMult and RotatE. The embeddings for FB15K-237 were trained using the entire dataset, through pykeen's hyperparameter optimization pipeline. For Wikidata5m, we use pre-trained embeddings from GraphVite[2] (Zhu et al., 2019). To measure DPD and PPD, a random forest classifier was trained on the task of profession prediction in both datasets. We attempt classification of the 5 and 10 most common professions in FB15K-237 and Wikidata5m respectively, and relabel the rest as "OTHER". A pre-processing step was applied to remove any tails that appeared less than 10 times in the test set.

4.1 Results

Table 1 compares all three bias measurements of the three most common relations in the FB15k-237 dataset. We observe that across all embeddings, the relation *gender* has the lowest DPD bias, and *nationality* the highest bias in both DPD and PPD, suggesting it may need debiasing. Moreover, no PPD bias is detected for *languages*. The common patterns across embeddings might imply that the biases do not arise from the embedding methods, but are rather inherent to the data itself or to the classifier. TLB presents a more mixed picture, with the most biased relation varying between embeddings. Since TLB is calculated using the score function of the embedding model, it is likely to be more sensitive to the KGE method.

The aggregated DPD and PPD bias scores on Wikidata5m are shown in Table 2. The relation portraying highest DPD on this dataset by a margin is *position played on team/specialty*, followed by *sport*. While our framework would mark these two relations as biased according to DPD, the practitioner might choose not to debias them, since they are related to a person's profession. Notably the PPD for these two relations is low, further illustrating the importance of offering a multi-bias perspective for a more robust bias evaluation. On the other hand, *given name* scores relatively high on both DPD and PPD in most embeddings, and can be considered an unwanted bias by the practitioner, since a person's given name should normally not affect their occupation. Therefore, given these scores, one may choose to only debias the relation *given name* in Wikidata5m.

Lastly, we provide a qualitative example shown in Table 3, presenting the disaggregated TLB bias scores with respect to *nationality* in FB15k-237. We display the five professions with highest TLB with respect to England versus the United States, the two most common tails for the relation *nationality*. At the fine-grained level, we notice a historical stereotype might remain, where England associates more with scientific occupations while the U.S. is biased towards entertainment careers. Moreover, the bias towards England appears lower, namely, the highest TLB bias towards England is significantly lower than the highest bias towards the U.S. The disagreggated tables presenting TLB bias for the other embedding methods and relations can be found in the Appendix A.3.

5 Discussion

In this paper, we proposed a novel framework to systematically identify, measure and inform biases in knowledge graph embeddings (KGE). The contribution of our model is to aid stakeholders and practitioners with a quantitative approach to identify biased relations in the KGE. Since biases are context- and culture-dependent, the final determination on what to debias may depend on the downstream task and is left to the practitioner. For example, one would want to remove gender biases from a question answering task about historical figures, while in medical related data, keeping gender information can be valuable for proper diagnosis.

Our implementation provides the user with bias scores rather than a binary decision. The choice of which relations to debias is then to be done by comparing the relative scores of relations in the KG, combined with domain knowledge. In future work, we would like to derive a threshold that can provide users with a binary score (biased/unbiased) for each

[2] the embeddings can be found at https://graphvite.io/docs/latest/pretrained_model.html

	Demographic Parity Distance				Predictive Parity Distance			
	TransE	ComplEx	RotatE	DistMult	TransE	ComplEx	RotatE	DistMult
country of citizenship	0.53	0.55	0.54	0.52	0.07	0.08	0.09	0.07
given name	0.59	0.63	0.51	0.60	**0.1**	0.11	0.11	0.09
place of birth	0.52	0.51	0.47	0.51	0.06	0.08	0.02	0.06
sport	0.73	0.74	0.78	0.64	0.0	0.0	0.0	0.0
languages spoken	0.46	0.57	0.49	0.54	0.07	0.09	**0.14**	0.08
position played on team / speciality	**1.21**	**1.26**	**1.24**	**1.11**	0.06	**0.14**	**0.14**	**0.12**

Table 2: Bias scores for most the common relations in Wikidata5m under a profession prediction task. The relation "position played on team/specialty" has the highest Demographic Parity Distance bias by a margin, which can be explained by its direct relation to profession. We further note that bias patterns are similar across embeddings.

England		U.S.	
Mathematician	0.0160	Television director	0.0250
Biologist	0.0158	Television producer	0.0227
Football player	0.0133	Screenwriter	0.0222
Physician	0.0105	Radio personality	0.0214
Scientist	0.0100	Actor	0.0207

Table 3: Professions with the highest Translational Likelihood Bias with respect to English versus U.S. nationalities in FB15K-237, using the TransE embedding.

relation, possibly through a statistical significance test. While a yes/no suggestion could save time and target a broader range of users, a careful analysis is required in order to define such a threshold without incurring further biases.

In summary, our paper presents a framework for quantifying bias in KGs, and by doing so identifies useful avenues for future research, and opens the possibility to compare various sources and definitions of bias. Janowicz et al. (2018) raise the concern that debiasing is not a neutral task, but rather based on social norms and is at risk of becoming censorship. By presenting a numerical method for selecting which relations to debias, we aim to minimize these risks. We hope to have illuminated the importance of identifying bias, as a complimentary component to algorithms that mitigate it.

References

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. 2020. Pykeen 1.0: A python library for training and evaluating knowledge graph embeddings. *CoRR*, abs/2007.14175.

Mario Arduini, Lorenzo Noci, Federico Pirovano, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. 2020. Adversarial learning for debiasing knowledge graph embeddings. *MLG Workshop at the 26th SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2020).*

Matthias Baumgartner, Wen Zhang, Bibek Paudel, Daniele Dell'Aglio, Huajun Chen, and Abraham Bernstein. 2018. Aligning knowledge base and document embedding models using regularized multi-task learning. In *International Semantic Web Conference*, pages 21–37. Springer.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Avishek Joey Bose and William Hamilton. 2019. Compositional fairness constraints for graph embeddings. *arXiv preprint arXiv:1905.10674*.

Joseph Fisher, Arpit Mittal, Dave Palfrey, and Christos Christodoulopoulos. 2020a. Debiasing knowledge graph embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7332–7345. Association for Computational Linguistics.

Joseph Fisher, Dave Palfrey, Christos Christodoulopoulos, and Arpit Mittal. 2020b. Measuring social bias in knowledge graph embeddings.

Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. Openke: An open toolkit for knowledge embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 139–144.

Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge graph embedding based question answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 105–113.

K. Janowicz, Bo Yan, Blake Regalia, R. Zhu, and Gengchen Mai. 2018. Debiasing knowledge graphs: Why female presidents are not like female popes. In *International Semantic Web Conference*.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores.

Nancy Krieger, Pamela Waterman, Jarvis Chen, Mah-J Soobader, S Subramanian, and Rosa Carson. 2002. Zip code caveat: Bias due to spatiotemporal mismatches between zip codes and us census–defined geographic areas—the public health disparities geocoding project. *American journal of public health*, 92:1100–2.

Chan Liu, Lun Li, Xiaolu Yao, and Lin Tang. 2019. A survey of recommendation algorithms based on knowledge graph embedding. In *2019 IEEE International Conference on Computer Science and Educational Informatization (CSEI)*, pages 168–171. IEEE.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163.

Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 43–54. Association for Computational Linguistics.

Kailash Karthik Saravanakumar. 2020. The impossibility theorem of machine fairness - A causal perspective. *CoRR*, abs/2007.06024.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. FairWare '18, page 1–7, New York, NY, USA. Association for Computing Machinery.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2019. Kepler: A unified model for knowledge embedding and pretrained language representation. *arXiv preprint arXiv:1911.06136*.

Tim Brennan William Dieterich, Christina Mendoza. 2016. Compas risk scales : Demonstrating accuracy equity and predictive parity.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics.

Zhaocheng Zhu, Shizhen Xu, Meng Qu, and Jian Tang. 2019. Graphvite: A high-performance cpu-gpu hybrid system for node embedding. In *The World Wide Web Conference*, pages 2494–2504. ACM.

A Appendix

A.1 Profession Classifier

We trained a random forest and an MLP classifier to predict the occupations on the KGEs. On random forest, we did hyperparameter search on maximal depths in $[3, 4, 5, 6]$ and batch sizes in $[100, 256, 500]$. On FB15K-237, we chose the random forest classifier with maximal depth of 4 and balanced class weights and a batch size of 256 as our final model, as it had the best performance. On Wikidata5m we choose the MLP classifier. The accuracy and balanced accuracy on classifying each entity into 6 and 11 occupation classes on FB15K-237 and Wikidata5M respectively are presented in tables 4 and 5.

	TransE	ComplEx	DistMult	RotatE
accuracy	0.5	0.514	0.499	0.517
balanced accuracy	0.329	0.34	0.33	0.356

Table 4: Performance of the random forest classifier on a 6 class classification task, predicting occupation on FB15K-237.

	TransE	ComplEx	DistMult	RotatE
accuracy	0.7	0.67	0.68	0.63
balanced accuracy	0.61	0.55	0.55	0.44

Table 5: Performance of the MLP classifier on a 11 class classification task, predicting occupation on Wikidata5m.

A.2 Knowledge Graph Embeddings

For the purpose of this paper, we trained a range of knowledge graph embedding models on FB15K237. The hits@k scores of the embeddings are listed in Table 6 below. We trained the embeddings through the hyperparameter optimization pipeline of pykeen (Ali et al., 2020), or by using the suggested parameters either from pykeen or openKE (Han et al., 2018).

	TransE	ConvE	ComplEx	DistMult	RotatE
hits@10	0.42	0.308	0.183	0.366	0.446
hits@3	0.271	0.175	0.183	0.219	0.289
hits@1	0.094	0.184	0.183	0.118	0.175

Table 6: Hit@k for the trained embeddings.

A.3 Translational Likelihood scores on FB15K-237

Below we present the disaggregated Translational Likelihood Bias (TLB) scores on FB15K-237, for the top three relations *nationality*, *language*, and *gender*.

A.3.1 TransE

In Table 7 and 8, we provide results using TransE embeddings on FB15k-237 dataset.

Male		Female	
Cinematographer	0.0367	Model	0.0157
Farmer	0.0365	Pin-up model	0.0131
Soldier	0.0346	Spokesperson	0.0077
/m/0196pc	0.0295	VJ	0.0044
Screenwriter	0.0288	Environmentalist	0.0022

Table 7: Male v.s. Female.

English		Hindi	
Model	0.0213	Stunt performer	0.0060
Author	0.0205	Music Director	0.0054
Singer-songwriter	0.0199	Prime Minister of Canada	0.0008
Designer	0.0199	Politician	0.0002
Spokesperson	0.0189	Storyboard artist	-0.0008

Table 8: English language v.s. Hindi language.

A.3.2 ComplEx

In Table 9, 10 and 11, we provide results using ComplEx embeddings on FB15k-237 dataset.

Male		Female	
Football Player	0.0006	Television Producer	0.0015
Politician	0.0003	Comedian	0.0014
Lawyer	0.0002	Prime Minister of Canada	0.0014
Architect	0.0002	Television director	0.0013
Mathematician	0.0002	Dub Actor	0.0012

Table 9: Male v.s. Female.

English		Hindi	
Make-up artist	-0.0021	Theatrical producer	0.0046
Production sound mixer	-0.0021	Supermodel	0.0046
Art Director	-0.0024	Music video director	0.0045
/m/089fss	-0.0025	VJ	0.0045
Football player	-0.0025	Pin-up model	0.0045

Table 10: English language v.s. Hindi language.

England		U.S.	
Architect	0.0047	Production sound mixer	-0.0012
Mathematician	0.0046	Make-up artist	-0.0012
Scientist	0.0046	Voice Actor	-0.0017
Critic	0.0046	Art Director	-0.0017
Inventor	0.0046	Television producer	-0.0018

Table 11: England v.s. U.S..

A.3.3 DistMult

In Table 12, 13 and 14, we provide results using DistMult embeddings on FB15k-237 dataset.

Male		Female	
/m/0196pc	0.0054	Pin-up model	0.0067
Cinematographer	0.005	Model	0.0059
Soldier	0.0049	Supermodel	0.0043
/m/01c8w0	0.0049	/m/064xm0	0.0026
Mathematician	0.0046	Prime Minister of Canada	0.0024

Table 12: Male v.s. Female.

English		Hindi	
Author	0.0021	/m/028kk_	0.0005
Artist	0.002	Costume designer	0.0005
Actor	0.0019	Audio engineer	0.0002
/m/0np9r	0.0019	Cinematographer	0.0002
Spokesperson	0.0019	Prime Minister of Canada	0.0002

Table 13: English language v.s. Hindi language.

England		U.S.	
Physician	0.0013	/m/0196pc	0.0024
Mathematician	0.0013	Screenwriter	0.0022
Scientist	0.001	Radio personality	0.0022
/m/0q04f	0.0009	/m/02krf9	0.0021
Football player	0.0008	Animator	0.002

Table 14: England v.s. U.S.

A.3.4 RotatE

In Tables 15, 16 and 17, we provide results using RotatE embeddings on FB15k-237 dataset.

Male		Female	
/m/0196pc	0.0002	Model	0.0002
Cinematographer	0.0002	Pin-up model	0.0002
Inventor	0.0002	Supermodel	0.0002
/m/01c8w0	0.0002	Spokesperson	0.0001
Composer	0.0002	VJ	0.0001

Table 15: Top 5 biased professions in terms of gender.

English		Hindi	
Theatrical producer	0.0002	/m/01tkqy	0.0
Spokesperson	0.0002	Politician	0.0
Author	0.0002	Prime Minister of Canada	0.0
Musician	0.0002	/m/028kk_	0.0
Singer-songwriter	0.0002	Football player	-0.0001

Table 16: Top 5 biased professions: English language v.s. Hindi language.

England		U.S.	
Biologist	0.0001	Attorneys in the United States	0.0002
Mathematician	0.0001	/m/0196pc	0.0002
Football player	0.0001	Music executive	0.0002
Physician	0.0001	Television producer	0.0002
/m/0q04f	0.0001	Businessperson	0.0002

Table 17: Top 5 biased professions.