Evaluating Gender Bias in Hindi-English Machine Translation

Gauri Gupta*
Manipal Institute of Technology
MAHE, Manipal, 576104
gaurigupta.315@gmail.com

Krithika Ramesh*
Manipal Institute of Technology
MAHE, Manipal, 576104
kramesh.tlw@gmail.com

Sanjay Singh
Manipal Institute of Technology
MAHE, Manipal, 576104
sanjay.singh@manipal.edu

Abstract

With language models being deployed increasingly in the real world, it is essential to address the issue of the fairness of their outputs. The word embedding representations of these language models often implicitly draw unwanted associations that form a social bias within the model. The nature of gendered languages like Hindi, poses an additional problem to the quantification and mitigation of bias, owing to the change in the form of the words in the sentence, based on the gender of the subject. Additionally, there is sparse work done in the realm of measuring and debiasing systems for Indic languages. In our work, we attempt to evaluate and quantify the gender bias within a Hindi-English machine translation system. We implement a modified version of the existing TGBI metric based on the grammatical considerations for Hindi. We also compare and contrast the resulting bias measurements across multiple metrics for pre-trained embeddings and the ones learned by our machine translation model.

1 Introduction

There has been a recent increase in the studies on gender bias in natural language processing considering bias in word embeddings, bias amplification, and methods to evaluate bias (Savoldi et al., 2021), with some evaluation methods introduced primarily to measure gender bias in MT systems. In MT systems, bias can be identified as the cause of the translation of gender-neutral sentences into gendered ones. There has been little work done for bias in language models for Hindi, and to the best of our knowledge, there has been no previous work that measures and analyses bias for MT of Hindi. Our approach uses two existing and broad frameworks for assessing bias in MT, including the Word Embedding Fairness Evaluation (Badilla et al., 2020) and the Translation Gender Bias Index (Cho et al., 2019) on Hindi-English MT systems. We modify some of the existing procedures within these metrics required for compatibility with Hindi grammar. This paper contains the following contributions:

1. Construction of an equity evaluation corpus (EEC) (Kiritchenko and Mohammad, 2018) for Hindi of size 26370 utterances using 1558 sentiment words and 1100 occupations following the guidelines laid out in Cho et al. (2019).

2. Evaluation of gender bias in MT systems for Indic languages.

3. An emphasis on a shift towards inclusive models and metrics. The paper is also demonstrative of language that should be used in NLP papers working on gender bias.

All our codes and files are publicly available.[1]

2 Related Work

The prevalence of social bias within a language model is caused by it inadvertently drawing unwanted associations within the data. Previous works that have addressed tackling bias include Bolukbasi et al. (2016), which involved the use of multiple gender-definition pairs and principal component analysis to infer the direction of the bias. In order to mitigate the bias, each word vector had its projection on this subspace subtracted from it. However, this does not entirely debias the word vectors, as noted in Gonen and Goldberg (2019).

[1] https://github.com/stolenpyjak/hi-en-bias-eval

Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, pages 16–23
August 5, 2021. ©2021 Association for Computational Linguistics

There have been various attempts to measure the bias in existing language models. Huang et al. (2020) measure bias based on whether the sentiment of the generated text would alter if there were a change in entities such as the occupation, gender, etc. Kurita et al. (2019) performed experiments on evaluating the bias in BERT using the Word Embedding Association Test (WEAT) as a baseline for their own metric, which involved calculating the mean of the log probability bias score for each attribute.

Concerning the measurement of bias in existing MT systems, Stanovsky et al. (2019) came up with a method to evaluate gender bias for 8 target languages automatically. Their experiments aligned translated text with the source text and then mapped the English entity (source) to the corresponding target translation, from which the gender is extracted.

Most of the focus in mitigating bias has been in English, which is not a gendered language. Languages like Hindi and Spanish contain grammatical gender, where the gender of the verbs, articles, adjectives must remain consistent with that of the gender of the noun. In Zhou et al. (2019) a modified version of WEAT was used to measure the bias in Spanish and French, based on whether the noun was inanimate or animate, with the latter containing words like 'doctor,' which have two variants for 'male' and 'female' each. Gonen et al. (2019) worked on addressing the problem with such inanimate nouns as well and attempted to neutralize the grammatical gender signal of these words during training by lemmatizing the context words and changing the gender of these words.

While there has been much work on quantifying and mitigating bias in many languages in NLP, the same cannot be said for Hindi and other Indic languages, possibly because they are low-resource. Pujari et al. (2019) was the first work in this area; they use geometric debiasing, where a bias subspace is first defined and the word is decomposed into two components, of which the gendered component is reduced. Finally, SVMs were used to classify the words and quantify the bias.

3 Methodology

3.1 Dataset and Data Preprocessing

The trained model that we borrowed from Gangar et al. (2021) was trained on the IIT-Bombay Hindi-English parallel data corpus (Kunchukuttan et al., 2018), which contains approximately 1.5 million examples across multiple topics. Gangar et al. (2021) used back-translation to increase the performance of the existing model by training the English-Hindi model on the IIT-Bombay corpus and then subsequently used it to translate 3 million records in the WMT-14 English monolingual dataset to augment the existing parallel corpus training data. The model was trained on this back-translated data, which was split into 4 batches.

The dataset cleaning involved removing special characters, punctuation, and other noise, and the text was subsequently converted to lowercase. Any duplicate records within the corpus were also removed, word-level tokenization was implemented, and the most frequent 50,000 tokens were retained. In the subword level tokenization, where byte-pair encoding was implemented, 50,000 subword tokens were created and added to this vocabulary.

3.2 NMT Model Architecture

For our experiments in building the neural machine translation model, we made use of the OpenNMT-tf (Klein et al., 2020) library, with the model's configuration being borrowed from Gangar et al. (2021). The OpenNMT model made use of the Transformer architecture (Vaswani et al., 2017), consisting of 6 layers each in the encoder and decoder architecture, with 512 hidden units in every hidden layer. The dimension of the embedding layer was set to 512, with 8 attention heads, with the LazyAdam optimizer being used to optimize model parameters. The batch size was 64 samples, and the effective batch size for each step was 384.

3.3 WEFE

The Word Embedding Fairness Evaluation framework is used to rank word embeddings using a set of fairness criteria. WEFE takes in a query, which is a pair of two sets of target words and sets of attribute words each, which are generally assumed to be characteristics related to the former.

$$Q = (\{T_{women}, T_{men}\}, \{A_{career}, A_{family}\}) \quad (1)$$

The WEFE ranking process takes in an input of a set of multiple queries which serve as tests across which bias is measured Q, a set of pre-trained word embeddings M, and a set of fairness metrics F.

3.3.1 The Score Matrix

Assume a fairness metric K is chosen from the set F, with a query template $s = (t, a)$, where all

Embedding	WEAT	RNSB	RND	ECT
NMT-English-(512D)	0.326529	0.018593	0.065842	0.540832
w2v-google-news-300	0.638202	0.01683	0.107376	0.743634
hi-300	0.273154	0.02065	0.168989	0.844888
NMT-Hindi-(512D)	0.182402	0.033457	0.031325	0.299023

Table 1: This table depicts the results for the various metrics that were used on the embeddings, and the final values based on their ranking by the Word Embedding Fairness Evaluation Framework.

subqueries must satisfy this template. Then,

$$Q_K = Q_1(s) \cup Q_2(s) \cup ... \cup Q_r(s) \qquad (2)$$

In that case, the $Q_i(s)$ forms the set of all subqueries that satisfy the query template. Thus, the value of $F = (m, Q)$ is computed for every pretrained embedding m that belongs to the set M, for each query present in the set. The matrix produced after doing this for each embedding is of the dimensions $M \times Q_K$.

The rankings are created by aggregating the scores for each row in the aforementioned matrix, which corresponds to each embedding. The aggregation function chosen must be consistent with the fairness metric, where the following property must be satisfied for $\leq_F$, where $x, x^{'}, y, y^{'}$ are random values in $\rm I\!R$, then $agg(x, x^{'}) \leq agg(y, y^{'})$ must hold true to be able to use the aggregation function. The result after performing this operation for every row is a vector of dimensions $1 \times M$, and we use $\leq F$ to create a ranking for every embedding, with a smaller score being ranked higher than lower ones.

After performing this process for every fairness metric over each embedding $m \in M$, the resultant matrix with dimensions $M \times F$ consisting of the ranking indices of every embedding for every metric, and this allows us to compare and analyze the correlations of the different metrics for every word embedding.

3.4 Metrics

3.4.1 WEAT

The WEAT (Word Embedding Association Test) (Caliskan et al., 2017) metric, inspired by the IAT (Implicit Association Test), takes in a set of queries as its input, with the queries consisting of sets of target words, and attribute words. In our case, we have defined two sets of target words catering to the masculine and feminine gendered words, respectively. In addition to this, we have defined multiple pairs of sets of attribute words, as mentioned in the Appendix. WEAT calculates the association of the target set T_1 with the attribute set A_1 over the attribute set A_2, relative to T_2. For example, as observed in Table 1, the masculine words tend to have a greater association with career than family than the feminine words. Thus, given a word w in the word embedding:

$$\text{d(w,A}_1, A_2) = (mean_{x \in A_1} cos(w, x)) - (mean_{x \in A_2} cos(w, x)) \qquad (3)$$

The difference of the mean of the cosine similarities of a given word's embedding vector with the word embedding vectors of the attribute sets are utilized in the following equation to give an estimate of the association.

$$\text{F}_{WEAT}(M, Q) = \Sigma_{w \in T_1} d(w, A_1, A_2) - \Sigma_{w \in T_2} d(w, A_1, A_2) \qquad (4)$$

3.4.2 RND

The objective of the Relative Norm Distance (RND) (Garg et al., 2018) is to average the embedding vectors within the target set T, and for every attribute $a \in A$, the norm of the difference between the average target and the attribute word is calculated, and subsequently subtracted.

$$\sum_{x \in A} (\|avg(T_1) - x\|_2 - \|avg(T_2) - x\|_2) \qquad (5)$$

The higher the value of the relative distance from the norm, the more associated the attributes are with the second target group, and vice versa.

3.4.3 RNSB

The Relative Negative Sentiment Bias (RNSB) (Sweeney and Najafian, 2019) takes in multiple target sets and two attribute sets and creates a query. Initially, a binary classifier is constructed, using the first attribute set A_1 as training examples for the first class, and A_2 for the second class. The classifier subsequently assigns every word w a probability, which implies its association with an attribute set, i.e

$$p(A_1) = C_{(A_1, A_2)}(w) \qquad (6)$$

Here, $C_{(A_1,A_2)}(x)$ represents the binary classifier for any word x. The probability of the word's association with the attribute set A_2 would therefore be calculated as $1 - C_{(A_1,A_2)}(w)$. A probability distribution P is formed for every word in each of the target sets by computing this degree of association. Ideally, a uniform probability distribution U should be formed, which would indicate that there is no bias in the word embeddings with respect to the two attributes selected. The less uniform the distribution is, the more the bias. We calculate the RNSB by defining the Kulback-Leibler divergence of P from U to assess the similarity of these distributions.

3.4.4 ECT

The Embedding Coherence Test (Dev and Phillips, 2019) compares the vectors of the two target sets T_1 and T_2, averaged over all their terms, with vectors from an attribute set A. It does so by computing mean vectors for each of these target sets such that:

$$\mu_i = \frac{1}{|T_i|}\Sigma_{t_i \in T_i}\, t_i \tag{7}$$

After calculating the mean vectors for each target set, we compute its cosine similarity with every attribute vector $a \in A$, resulting in s_1 and s_2, which are vector representations of the similarity score for the target sets. The ECT score is computed by calculating the Spearman's rank correlation between the rank orders of s_1 and s_2, with a higher correlation implying lower bias.

3.5 TGBI

The Translation Gender Bias Index (TGBI) is a measure to detect and evaluate the gender bias in MT systems, introduced by Cho et al. (2019). They use Korean-English (KN-EN) translation. In Cho et al. (2019), the authors create a test set of words or phrases that are gender neutral in the source language, Korean. These lists were then translated using three different models and evaluated for bias using their evaluation scheme. The evaluation methodology proposed in the paper quantifies associations of 'he,' 'she,' and related gendered words present translated text. We carry out this methodology for Hindi, a gendered low-resource language in natural language processing tasks.

3.5.1 Occupation and Sentiment Lists

Considering all of the requirements laid out by Cho et al. (2019), we created a list of unique occupations and positive and negative sentiment in our source language, Hindi. The occupation list was generated by translating the list in the original paper. The translated lists were manually checked for errors and for the removal of any spelling, grammatical errors, and gender associations within these lists by native Hindi speakers. The sentiment lists were generated using the translation of existing English sentiment lists (Liu et al., 2005; Hu and Liu, 2004) and then manually checked for errors by the authors. This method of generation of sentiment lists in Hindi using translation was also seen in Bakliwal et al. (2012).

The total lists of unique occupations and positive and negative sentiment words come out to be 1100, 820 and 738 in size respectively. These lists have also been made available online.[2]

3.5.2 Pronouns and Suffixes

Hindi, unlike Korean, does not have gender-specific pronouns in the third person. Cho et al. (2019) considered 그 사람 (ku salam), 'the person' as a formal gender-neutral pronoun and the informal gender-neutral pronoun, 걔 (kyay) for a part of their gender-neutral corpus. However, for Hindi, we directly use the third person gender-neutral pronouns. This includes वह (vah), वे (ve), वो (vo) corresponding to formal impolite (familiar), formal polite (honorary) and informal (colloquial) respectively (Jain, 1969).

As demonstrated by Cho et al. (2019), the performance of the MT system would be best evaluated with different sentence sets used as input. We apply the three categories of Hindi pronouns to make three sentence sets for each lexicon set (sentiment and occupations): (i) formal polite, (ii) formal impolite, and (iii) informal (colloquial use).

3.5.3 Evaluation

We evaluate two systems, Google Translate and the Hi-En OpenNMT model, for seven lists that include: (a) informal, (b) formal, (c) impolite, (d) polite, (e) negative, (f) positive, and (g) occupation that are gender-neutral. We have attempted to find bias that exists in different types of contexts using these lists. The individual and cumulative scores help us assess contextual bias and overall bias in Hi-En translation respectively.

TGBI uses the number of translated sentences that contain she, he or they pronouns (and conventionally associated[3] words such as girl, boy or

[2] https://github.com/stolenpyjak/hi-en-bias-eval

[3] The distinction between pronouns, gender and sex has

Sentence	Size	OpenNMT-tf	Google Translate
Informal	2628	0.7543 (0.0315, 0.7473)	0.3553 (0.2763, 0.2146)
Formal	5286	0.5410 (0.0773, 0.5090)	0.5464 (0.1015, 0.5066)
Impolite	2628	0.2127 (0.1552, 0.0966)	0.2716 (0.1990, 0.1400)
Polite	2658	0.9168 (0.0003, 0.9168)	0.8690 (0.0052 0.8683)
Positive	2460	0.6765 (0.0825, 0.6548)	0.5819 (0.1589, 0.5329)
Negative	2212	0.6773 (0.0641, 0.6773)	0.5384 (0.15822, 0.5384)
Occupation	3242	0.5100 (0.0453, 0.4888)	0.3599 (0.1610, 0.2680)
Average:		**0.6127**	**0.5032**

Table 2: The values present under each MT system shows it's corresponding $P_i(p_{she}, p_{they})$ value for each sentence set and the average TGBI value is calculated in the last row.

person) to measure bias by associating that pronoun with p_{he}, p_{she} and p_{they}[4] for the scores of P_1 to P_7 corresponding to seven sets S_1 to S_7 such that:

$$P_i = \sqrt{(p_{he} * p_{she} + p_{they})} \qquad (8)$$

and finally, TGBI = avg(P_i).

4 Results and Discussion

The BLEU score of the OpenNMT model we used was 24.53, and the RIBES score was 0.7357 across 2478 samples.

4.1 WEAT

We created multiple sets of categories for the attributes associated with 'masculine' and 'feminine,' including the subqueries as listed in the supplementary material. We used both the embeddings from the encoder and the decoder, that is to say, the source and the target embeddings, as the input to WEFE alongside the set of words defined in the target and attribute sets. Aside from this, we have also tested pre-trained word embeddings that were available with the gensim (Rehurek and Sojka, 2011) package on the same embeddings. The results of the measurement of bias using the WEFE framework are listed in Table 1.

For the English embeddings, there is a significant disparity in the WEAT measurement for the Math vs Arts and the Science vs Arts categories. This could be owing to the fact that there is little data in the corpus that the MT system was trained over, which is relevant to the attributes in these sets. Hence the bias is minimal compared to the pre-trained word2vec embeddings, which is learned over a dataset containing 100 billion words and is likely to learn more social bias compared to the embeddings learned in the training of the MT system. We notice a skew in some of the other results, which could be due to the MT model picking up on gender signals that have strong associations of the target set with the attribute set, implying a strong bias in the target set training data samples itself. However, all of these metrics and the pre-trained embeddings used are in positive agreement with each other regarding the inclination of the bias.

For the Hindi embeddings, while the values agree with each other for the first two metrics, there is a much more noticeable skew in the RND and ECT metrics. The pre-trained embeddings seem to exhibit much more bias, but the estimation of bias within the embedding learned by the MT may not be accurate due to the corresponding word vectors not containing as much information, consider the low frequency of terms in the initial corpus that the NMT was trained on. In addition to this, there were several words in the attribute sets in English that did not have an equivalent Hindi translation or produced multiple identical attribute words in Hindi. Consequently, we had to modify the Hindi attribute lists.

While these metrics can be used to quantify gender bias, despite not necessarily being robust, as is illustrated in Ethayarajh et al. (2019) which delves into the flaws of WEAT, they also treat gender in binary terms, which is also a consistent trend across research related to the field.

Our findings show a heavy tendency for Hi-En MT systems to produce gendered outputs when the gender-neutral equivalent is expected. We see that many stereotypical biases are present in the source and target embeddings used in our MT system. Further work to debias such models is necessary, and the development of a more advanced NMT would

been explain in section 5.2

[4] Changed convention to disassociate pronouns with gender and sex

be beneficial to produce more accurate translations to be studied for bias.

4.2 TGBI

The final TGBI score which is the average of different P_i values, is between 0 and 1. A score of 0 corresponds to high bias (or gendered associations in translated text) and 1 corresponds to low bias (Cho et al., 2019).

The bias values tabulated in Table 2, show that within both models, compared to the results on sentiment lexicons, occupations show a greater bias, with p_{she} value being low. This points us directly to social biases projected on the lexicons (S_{bias}[5]). For politeness and impoliteness, we see that the former has the least bias and the latter most across all lists. While considering formal and informal lists, informal pronoun lists show higher bias. There are a couple of things to consider within these results: a) the polite pronoun वे (ve) is most often used in plural use in modern text (V_{bias}), thus leading to a lesser measured bias, b) consider that both polite and impolite are included in formal which could correspond to its comparatively lower index value compared to informal.

Bias in MT outputs whether attributed to S_{bias} or V_{bias}, is harmful in the long run. Therefore, in our understanding, the best recommendation is that TGBI = 1 with corresponding p_{they}, p_{she}, p_{he} values 1, 0, 0 respectively.

5 Bias Statement

5.1 Bias Statement

In this paper, we examine gender bias in Hi-En MT comprehensively with different categories of occupations, sentiment words and other aspects. We consider bias as the stereotypical associations of words from these categories with gender or more specifically, gendered words. Based on the suggestions by Blodgett et al. (2020), we have the two main categories of harms generated by bias: 1) representational, 2) allocational. The observed biased underrepresentation of certain groups in areas such as Career and Math, and that of another group in Family and Art, causes direct representational harm. Due to these representational harms in MT and other downstream applications, people who already belong to systematically marginalized groups are put further at risk of being negatively affected by stereotypes. Inevitably, gender bias causes errors in translation (Stanovsky et al., 2019) which can contribute to allocational harms due to disparity in how useful the system proves to be for different people, as described in an example in Savoldi et al. (2021). The applications that MT systems are used to augment or directly develop increase the risks associated with these harms.

There is still only a very small percent of the second most populated country in the world, India that speaks English, while English is the most used language on the internet. It is inevitable that a lot of content that might be consumed now or in the future might be translated. It becomes imperative to evaluate and mitigate the bias within MT systems concerning all Indic languages.

5.2 Ethical Considerations and Suggestions

There has been a powerful shift towards ethics within the NLP community in recent years and plenty of work in bias focusing on gender. However, we do not see in most of these works a critical understanding of what gender means. It has often been used interchangeably with the terms 'female' and 'male' that refer to sex or the external anatomy of a person. Most computational studies on gender see it strictly as a binary, and do not account for the difference between gender and sex. Scholars in gender theory define gender as a social construct or a learned association. Not accommodating for this definition in computational studies not only oversimplifies gender but also possibly furthers stereotypes (Brooke, 2019). It is also important to note here that pronouns in computational studies have been used to identify gender, and while he and she pronouns in English do have a gender association, pronouns are essentially a replacement for nouns. A person's pronouns, like their name, are a form of self-identity, especially for people whose gender identity falls outside of the gender binary (Zimman, 2019). We believe research specifically working towards making language models fair and ethically sound should be employing language neutralization whenever possible and necessary and efforts to make existing or future methodologies more inclusive. This reduces further stereotyping (Harris et al., 2017; Tavits and Pérez, 2019). Reinforcing gender binary or the association of pronouns with gender may be invalidating for people who identify themselves outside of the gender binary (Zimman,

[5] In Cho et al. (2019), the authors describe two kinds of bias: V_{bias} which is based on the volume of appearance in the corpora and S_{bias} which is based on social bias that is projected in the lexicons.

2019).

6 Conclusion and Future Work

In this work, we have attempted to gauge the degree of gender bias in a Hi-En MT system. We quantify gender bias (so far only for the gender binary) by using metrics that take data in the form of queries and employ slight modifications to TGBI to extend it to Hindi. We believe it could pave the way to the comprehensive evaluation of bias across other Indic and/or gendered languages. Through this work, we are looking forward to developing a method to debias such systems and developing a metric to measure gender bias without treating it as an immutable binary concept.

7 Acknowledgements

The authors of the paper are grateful for the contributions of Rianna Lobo for their reviews on the Bias Statement and Ethics Section. The efforts of the reviewers in reviewing the manuscript and their valuable inputs are appreciated. We would also like to thank the Research Society MIT for supporting the project.

References

Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2020. Wefe: The word embeddings fairness evaluation framework. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 430–436. International Joint Conferences on Artificial Intelligence Organization. Main track.

Akshat Bakliwal, Piyush Arora, and Vasudeva Varma. 2012. Hindi subjective lexicon: A lexical resource for Hindi adjective polarity classification. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1189–1196, Istanbul, Turkey. European Language Resources Association (ELRA).

Su Lin Blodgett, Solon Barocas, Hal Daumé III au2, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520.

Sian Brooke. 2019. "condescending, rude, assholes": Framing gender and hostility on Stack Overflow. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 172–180, Florence, Italy. Association for Computational Linguistics.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Sunipa Dev and Jeff M. Phillips. 2019. Attenuating bias in word vectors. *CoRR*, abs/1901.07656.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Kavit Gangar, Hardik Ruparel, and Shreyas Lele. 2021. Hindi to english: Transformer-based neural machine translation. In *International Conference on Communication, Computing and Electronics Systems*, pages 337–347, Singapore. Springer Singapore.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages?

Chelsea A. Harris, Natalie Blencowe, and Dana A. Telem. 2017. What is in a pronoun? why gender-fair language matters. *Annals of surgery*, 266(6):932–933. 28902666[pmid].

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for*

Computational Linguistics: EMNLP 2020, pages 65–83, Online. Association for Computational Linguistics.

Dhanesh K. Jain. 1969. Verbalization of respect in hindi. *Anthropological Linguistics*, 11(3):79–97.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *CoRR*, abs/1805.04508.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The OpenNMT neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109, Virtual. Association for Machine Translation in the Americas.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, page 342–351, New York, NY, USA. Association for Computing Machinery.

Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2019. Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI 2019, page 450–456, New York, NY, USA. Association for Computing Machinery.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy. Association for Computational Linguistics.

Margit Tavits and Efrén O. Pérez. 2019. Language influences mass opinion toward gender and lgbt equality. *Proceedings of the National Academy of Sciences*, 116(34):16781–16786.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender.

Lal Zimman. 2019. Trans self-identification and the language of neoliberal selfhood: Agency, power, and the limits of monologic discourse. *International Journal of the Sociology of Language*, 2019:147–175.