Proceedings of LATERAISSE Workshop @LREC2022, pages 1–7
Marseille, 25 June 2022

Casteism in India, But not Racism - A Study of Bias in Word Embeddings of Indian Languages

Senthil Kumar B[1], Pranav Tiwari[2], Aman Chandra Kumar[2], Aravindan Chandrabose[1]
[1]Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
[2]Indian Institute of Information Technology, Tiruchirappalli, India
{senthil, aravindanc}@ssn.edu.in
{pranavtiwari548, amanchandrakumar202}@gmail.com

Abstract

In this paper, we studied the gender bias in monolingual word embeddings of two Indian languages Hindi and Tamil. Tamil is one of the classical languages of India from the Dravidian language family. In Indian society and culture, instead of racism, a similar type of discrimination called *casteism* is against the subgroup of peoples representing lower class or *Dalits*. The word embeddings measurement to evaluate bias using the WEAT score reveals that the embeddings are biased with gender and casteism which is in line with the common stereotypical human biases.

Keywords: bias in word embeddings, gender bias, caste bias, WEAT, Indian languages

1. Introduction

A language is a wonderful tool for communication. It has powered the human race for centuries and continues to be at the heart of our culture. India has more than 270 languages or dialects spoken as its mother tongue. Of 121 languages that are spoken by 10,000 or more people, 22 languages comprising 123 mother tongues are specified in the Eighth Schedule to the Constitution of India as Scheduled Languages. [1]. Hindi and Tamil are the Scheduled languages of India.

Based on cultural linkages and unfavorable social biases, NLP models are trained with a variety of biases and discrimination. Word embeddings have become a standard resource for representing the text in ML-based NLP applications. Generating a good word embedding is very important to avoid bias in the downstream tasks. Learning a high-quality word representation is extremely important for various syntactic and semantic tasks. The methods to evaluate the quality of word embeddings are categorized into intrinsic and extrinsic methods. Extrinsic methods use word embeddings as input features to a downstream task and measure changes in performance metrics specific to that task. But the intrinsic evaluation methods test the quality of an embedding independent of a specific NLP task. One technique to measure the quality of word embedding is to check whether it is unbiased towards gender, racism, religion, demographic, etc., using bias evaluation metrics like WEAT.

Despite the diversity, bias in word embeddings of Indian languages is studied less. So far, bias is experimented with Hindi, Bengali and Telugu languages of India. Hindi and Bengali are the languages of the Indo-Aryan (or Indic language) family. Tamil and Telugu are the languages of thfe Dravidian family and Tamil is a classical Dravidian language. Our study shows bias in the Tamil language which is highly agglutinate and also in Hindi. Instead of *racism*, we experiment with a type of bias called *casteism* which is highly prevalent in Indian culture. Caste systems in India have its root in medieval, early-modern, and modern India (Bayly, 2001).

The rest of the paper is structured as follows. Section 2 describes related works on these problems and provides context on why the problem is difficult and important to solve. Next, in sections 3, we describe the datasets and bias measure which are used to measure it. In Section 4 we analyse and present the results and conclusion about our work in section 5.

2. Related Work

Gender bias appears to be a common stereotype that exists across vast majority of data resources. An illustrious work by Bolukbasi et al. (2016) observed gender bias in Word2Vec word embeddings. They showed that gender bias could be found by identifying the direction in embedding subspace and could be neutralized. Caliskan et al. (2017) measured the bias in the Word2Vec embeddings on Google News corpus and pre-trained GloVe using WEAT, WEFAT score. Escudé Font et al. (2019) found gender bias in the translation of English-Spanish in the news domain. Embeddings

[1]Census of India, 2021

of gendered languages such as Spanish and French contain gender bias. Zhou et al. (2019) observed the bias in bilingual embeddings from MUSE while translating ES-EN and FR-EN, where both the Spanish and French are gendered languages. To neutralize gender in word embeddings, GN-GloVe (Zhao et al., 2018) is used to mitigate gender bias in word representations. Apart from gender bias, Manzini et al. (2019) found ethnicity and religion bias by extending WEAT to measure the bias over a Word2Vec model. Research on race in NLP remains less and ignored in many NLP tasks. Field et al. (2021) survey on racism in NLP research shows that only 13 papers from ACL anthology focus on racial bias in text representations (LMs, embeddings). The survey highlighted that the NLP research fails to account for the multidimensional race. Hasanuzzaman et al. (2017) shows that racism is in link with location information instead of gender. Bansal et al. (2021) measured gender bias using intrinsic, extrinsic bias and debias the word embeddings for three Indian languages (Hindi, Bengali, Telugu) in addition to English.
The challenges in Indian languages are:
1. The semantics of gender words may vary from one language to another.
2. While Bolukbasi et al. (2016) leverages the pronouns (e.g.,she/he) to construct gendered directions this might not be possible for many languages (e.g., In Tamil, the same pronoun **அவர்** is used to refer to both the male and female genders).
3. Certain terms in Tamil have male honorific forms, do not have the corresponding female honorific forms. One may be tempted to say the forms listed as masculine honorific forms are neutral forms. Yet, in actual use, these often assume male reference.

Male	Female	Honorific	English
பாடகன்	**பாடகி**	**பாடகர்**	singer
தலைவன்	**தலைவி**	**தலைவர்**	leader

Table 1: Gender-neutral or honorific terms in Tamil

2.1. Why Casteism but not Racism?

In gender classification based on photographs of faces, Buolamwini and Gebru (2018) could draw the connection between phenotype and race. They noted that racial categories are unstable and that phenotype can vary widely within a racial or ethnic category. Moreover Benthall and Haynes (2019) claims that the acquisition of a race by a person depends on several different factors, including bio-metric properties, socioeconomic class, and ancestral geographic and national origin. Hence Hanna et al. (2020; Benthall and Haynes (2019; Field et al. (2021) argue that race is a multi-dimensional and can refer to a variety of different perspectives. During the World Conference against Racism (WCAR) by United Nations in 2001, which discussed various manifestations of racism, the position of the Indian government was that the caste is not a race and hence is not relevant at conference (Pinto, 2001). Due to its multi-dimensional nature, no widely accepted categorization scheme and the Indian government stance, casteism is varied from racism.
Our contributions include considering two Indian languages, each from the Indo-Aryan and Dravidian families, and bias analysis concerning gender and casteism. As per the literature survey and to our knowledge, this is the first report on 1) bias in Tamil language embeddings, 2) the discrimination of subgroup of people in India under "casteism" is reflected in word embeddings. The choice of the current set of languages is motivated by the knowledge of the authors in these languages.

3. Experiment

Neural network models are quite powerful and efficient, but at the same time, these models inherently contain problematic biases in many forms. Many pre-trained language models such as Word2Vec, GloVe, ELMo, fastText, etc., are widely available for developers to generate word embeddings, but they should also be aware of what biases they contain and how they might exacerbate in those applications. In our experiment, two pre-trained language models: Word2Vec (Hindi) and fastText (Hindi and Tamil) are used to obtain the word embeddings. To check whether the embeddings of these models are biased or not, the WEAT metric is used to find its association or bias which is in line with the human bias.

3.1. Datasets

For gender bias, most of the words are taken from Caliskan et al. (2017) study on gender-biased words using male vs female and career vs family. The male vs female words is also measured against the male vs female traits (or adjectives). In Indian languages, some of the words are used in their transliterated form instead of their equivalent linguistic form. For example, the words उपचारिका (Nurse) is less frequently used instead of its transliterated form नर्स. The frequently used form is included in this study. Table 2 lists the statistics of the dataset for the Hindi and Tamil languages. Words such as loyal, family, happy, abuse, murder, assault and jail are taken from pleasant vs. unpleasant words of Caliskan et al. (2017). The other words are considered in the context of cultural and societal practices followed by the Indian people.

Targets	Hindi	Tamil	Attributes	Hindi	Tamil
Career vs Family	4	5	Male vs Female	10	7
Male vs. Female Traits	5	5	Male vs Female	10	11
Pleasant vs Unpleasant	18	8	Upper vs Lower	6	6
High-paid vs Low-paid	10	16	Upper vs Lower	6	6

Table 2: The number of words used in the target and attribute sets for Hindi and Tamil languages.

Gathering data to examine a new bias type called casteism in NLP is challenging. There is no exact translation of *caste* in Indian languages, but *varna* and *jati* are the two most approximate terms. The caste emanates from four *varnas* or *jati* system in Indian culture. For bias in casteism, the words are inferred from the four *varnas* system in India. The castes under four *varnas* or *jati* are grouped into a single, the remaining are considered as others or *untouchables* or Scheduled Castes, the official term as per the Constitution of India [2]. We label the group of four as upper and the other as lower caste. The peoples of upper caste are majority than the lower caste and hence lower caste is also referred to as minorities. The set of attribute words for caste in Hindi and Tamil is shown in Table 3 for upper caste and Table 4 for lower caste. '-' in the table indicates that a particular caste word is infrequently used in context in spite of its prevalence.

Hindi	Tamil	English
ब्राह्मण	பிராமணர்கள்	brahmins
क्षत्रिय	க்ஷத்திரியர்கள்	kshatriyas
वैश्य	வைசியர்கள்	vaisyas
-	சூத்திரர்கள்	kshudras
उच्च	உயர்	upper
पंडित	-	priest

Table 3: Hindi/Tamil upper caste words

Hindi	Tamil	English
हरिजन	ஹரிஜனங்கள்	harijans
दलितों	தலித்	dalits
अनुसूचित	அட்டவணைப்படுத்தப்பட்ட	schedule caste
अछूतों	தீண்டத்தகாதவர்கள்	untouchables
निचली	கீழ்	lower

Table 4: Hindi/Tamil lower caste words

3.2. Word2Vec model

Word2Vec [3] model trained on Hindi CoNLL 17 corpus using Continous Skipgram model in dimension 100.

[2]Caste System in India

[3]NLPL word embedding repository

3.3. fastText model

The fastText [4] is a pre-trained language model trained on Wikipedia and the Common Crawl to represent word vectors for different 157 languages. Each of these models was trained on Wikipedia dumps of the respective languages using CBOW with position-weights, in dimension 300, with character n-grams of length 5. It was observed that for languages with small Wikipedia, such as Finnish or Hindi, using the crawl data leads to great improvement in performance. However for the low resource languages such as Hindi, the quality of the obtained word vectors is much lower than for other languages (Grave et al., 2018).

3.4. Correlation with Human Biases using WEAT

We used the metric **Word Embedding Association Test (WEAT)** proposed by Caliskan et al. (2017) which uses permutation testing to demonstrate and quantify bias. WEAT measures the similarity of words by using the cosine between the pair of vectors of those words. It was applied to GloVe and Word2Vec vectors. WEAT can also be applied to other models. Consider the two sets of target words (like politician, engineer, tailor, ... and nanny, nurse, librarian, ...) and two sets of attribute words (like man, boy, ... and woman, girl ...) to measure the bias against the social attributes and roles. In mathematical terms, X and Y are assumed to be sets of target words of equal size, and A,B are the two sets of attribute words. The permutation test over X and Y is,

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

The degree of bias for each target concept is calculated as,

$$s(w, A, B) = mean_{a \in A} cos(\overrightarrow{w}, \overrightarrow{a}) - mean_{b \in B} cos(\overrightarrow{w}, \overrightarrow{b})$$

where $cos(\overrightarrow{a}, \overrightarrow{b})$ is the cosine similarity between the two vector embeddings. In other words, s(w,A,B) measures the association of the word w with the attribute, and s(X,Y,A,B) measures the

[4]fastText for different 157 languages

differential association of the two sets of target words with the attribute. The degree d to which the model associates the sets of target words with the sets of attribute words is,

$$d = \frac{mean_{x \in X} s(x, A, B) - mean_{y \in Y} s(y, A, B)}{std - dev_{w \in X \sqcup Y} s(w, A, B)}$$

For example, consider the target lists for the WEAT test are pleasant and unpleasant words, and the attributes are caste discrimination in India such as upper caste (e.g., "brahmins", "vaisyas", "kshatriyas") and lower caste (e.g., "dalits", "harijans", "untouchables"). The overall test score is the degree to which pleasant words are more associated with the upper caste, relative to lower caste. A high positive score means that pleasant words are more related to upper caste, and a high negative score means that unpleasant words are more associated with upper caste.

4. Result Analysis

Word2Vec (Skipgram) embeddings are used for Hindi language only. The fastText embeddings of Hindi and Tamil languages are measured for bias. We consider the following two target sets:
1) career vs. family, sentiment words or traits of male vs. female.
2) pleasant vs. unpleasant, career of upper vs. lower caste.
For the above target sets, the corresponding attribute sets are 1)male vs. female and 2)upper vs. lower caste. Note that from tables 5-12, the words are arranged in descending order of bias score. For example, occupation (career vs family) words are sorted with the degree of bias in descending order for Hindi in Table 5.

Male	Female
सेनाध्यक (commander)	बाई (maid)
सैनिक (soldier)	दाई (babysitter)
राजनीतिज्ञ (politician)	नर्स (nurse)
शिकारी (hunter)	रसोइया (cook)

Table 5: Hindi Male/Female-biased words for Occupation using fastText

For Tamil, the occupations of gender (career vs. family) differs from Hindi, because of the demographic or regional cultural influence. In both the languages, occupational words like politician, hunter and nurse, maid are biased towards male and female respectively.
The male vs female traits (adjectives) are different across the demography irrespective of gender as shown in Table 7 and 8. For example high degree of male trait word exercise (**உடற்பயிற்சி**) in Tamil

Male	Female
வேட்டைக்காரன் (hunter)	**பணிப்பெண்** (maid)
அரசியல்வாதி (politician)	**செவிலியர்** (nurse)
பொறியாளர் (engineer)	**ஒப்பனையாளர்** (stylist)
காவல் (police)	**நடனக்கலைஞர்** (dancer)
சிப்பாய் (soldier)	**கைவினை** (craft person)

Table 6: Tamil Male/Female-biased words for Occupation using fastText

is not the same in Hindi. For Hindi, it is combat (मुकाबल).

Male	Female
मुकाबला (combat)	सुंदरता (beauty)
अभ्यास (practice)	तलाक (divorce)
हमला (attack)	शादी (wedding)
घायल (injured)	परिपक्व (mature)
परिश्रम (hardwork)	प्यार (love)

Table 7: Hindi Male vs Female Traits (adjectives) using fastText

In both the languages, sentiment words like combat/battle, attack are associated towards male and beauty, wedding, divorce are associated towards female.

Male	Female
உடற்பயிற்சி (exercise)	**விவாகரத்து**(divorce)
இரக்கமற்ற (ruthless)	**அழகு** (beauty)
சக்தி (power)	**நகை** (jewel)
போர் (battle)	**திருமணம்**(wedding)
தாக்குதல் (attack)	**நளினம்** (elegance)

Table 8: Tamil Male vs Female Traits (adjectives) using fastText

4.1. Caste Bias in Indian Languages

Castes are rigid social groups characterized by hereditary transmission of lifestyle, occupation, and social status. This is ingrained in the social and economic status of peoples across castes in Indian culture. We measured the bias against the caste words for the two attribute sets: 1)Pleasant vs. unpleasant words and 2)Career words (high-paid vs low-paid). Some of the adjective words are used to denote a particular group of caste. Those words are categorized into pleasant and unpleasant words. The careers of the minority group or lower caste also differs from that of the upper caste group.

upper	lower
वैदिक (vaidic)	हमला (assault)
धनी (rich)	दुर्व्यवहार (abuse)
ज्ञान (knowledge)	जेल (jail)
भाग्यशाली (fortunate)	हत्या (murder)
निष्ठावान (loyal)	श्रम (labour)
साहित्य (literature)	निरक्षर (illiterate)
परिवार (family)	उत्पीड़ित (oppressed)
खुश (happy)	घृणा (hatred)
शक्ति (strength)	सताया (persecuted)

Table 9: Hindi Caste-biased Pleasant vs. unpleasant words using fastText

From the Table 9-10, the bias in the embeddings clearly shows the discrimination of the lower caste minority in India. India after 1947, enacted many affirmative action policies for the upliftment of historically marginalized groups. These policies included reserving a quota of places for these groups in higher education and government employment. But still, the word embeddings reflects the caste stereotypes that still exists in the Indian society. In Table 11-12, the bias in the embeddings clearly reflects the discrimination in the social-economic structure of the lower caste minority in India. The occupations of Dalits vary from caste to caste and geographical area. Most of them work with human waste, leather, dead bodies, etc., (Kaminsky; Long, 2011).

upper	lower
வேத (vedic)	**தாழ்த்தப்பட்ட** (downtrodden)
அறிவாளி (knowledge)	**ஒடுக்கப்பட்ட** (oppressed)
அதிர்ஷ்டசாலி(fortunate)	**அடிமைப்படுத்தப்பட்ட** (enslaved)
கல்வி (education)	**தாக்குதல்** (attack)
சக்தி (power)	**சிறை** (jail)
கற்றவர் (literate)	**கொலை** (murder)

Table 10: Tamil Caste-biased Pleasant vs. unpleasant words using fastText

Table-13 shows the WEAT scores for the different embedding models for the four different target and attribute sets. The score indicates that the direction of measured bias is in line with the common human biases. For the upper vs lower and career dataset, the negative WEAT score for the Word2Vec Hindi embeddings implies that the bias is against the common human biases. Generally, Hindi language embeddings are less biased than Tamil towards careers of upper and lower caste peoples. To prove that racism is not much prevalent in India, a set of racial prejudice words *chink, chinky, chinese, nepali* against the north-east Indians (Haokip, 2021) are paired with the pleasant vs. unpleasant words in Hindi. The negative score indicates that the embeddings are racial-free.

upper	lower
योद्धा (warrior)	मजदूरी (wage)
अफ़सर (officer)	बेरोज़गार (unemployed)
अभियंता (engineer)	कुम्हार *(potter)*
शिक्षक (teacher)	किसान *(farmer)*
वैज्ञानिक (scientist)	रक्षक *(protector)*
संगीत (music)	मोची *(cobbler)*
अनुसंधान (research)	चौकीदार *(watchman)*

Table 11: Hindi Caste-biased career words using fastText. Italicised is unbiased.

upper	lower
போர்வீரன் (warrior)	**கல்லறைத்தொழிலாளி** (cemetry worker)
வணிகர் (merchant)	**தொழிலாளி** (labour)
விஞ்ஞானி (scientist)	**துப்புரவாளர்** (sweeper)
பொறியாளர் (engineer)	**செருப்புத்தொழிலாளி** (cobbler)
அதிகாரி (officer)	**காவலாளி** (watchman)
ஆசிரியர் (teacher)	**விவசாயி** (farmer)

Table 12: Tamil Caste-biased career words using fastText

5. Conclusion

In this paper, instead of racism which is not applicable to India, casteism as per the the Indian social system is included in word embedding bias evaluation. We have identified the sets of caste words in Hindi and Tamil languages for caste bias analysis. WEAT metric is used to evaluate the word embeddings for gender and caste bias. The bias study on monolingual word embeddings of Word2Vec and fastText for two of the Indian languages such as Hindi and Tamil reveals that the gender and caste bias prevails in line with the stereotypes. From the literature and to our knowledge this is the first paper that reports the bias in Tamil word embeddings and caste bias in word embeddings of Indian languages. Also proved that the embeddings are racial-free.

In future, we will extend the bias analysis by including more Indian languages and apply debiasing techniques to mitigate the bias in Indian language word embeddings.

6. Bibliographical References

Bansal, S., Garimella, V., Suhane, A., and Mukherjee, A. (2021). Debiasing multilingual

Targets	Attributes	Word2Vec(H)	fastText(H)	fastText(T)
Career vs Family	Male vs Female	1.15	1.85	1.07
Male vs. Female Traits	Male vs Female	1.58	1.79	0.89
Pleasant vs Unpleasant	Upper vs Lower	1.08	1.52	1.84
High-paid vs Low-paid	Upper vs Lower	-0.38	0.99	1.55
Indian vs North-east	Pleasant vs Unpleasant	-	-0.36	-

Table 13: WEAT scores for different embedding models. Negative value indicates that the direction of the measured bias is against the common human biases. H-Hindi, T-Tamil

word embeddings: A case study of three indian languages. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, page 27–34, New York, NY, USA. Association for Computing Machinery.

Bayly, S. (2001). Cambridge University Press. ISBN 978-0-521-26434-1.

Benthall, S. and Haynes, B. D. (2019). Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 289–298, New York, NY, USA. Association for Computing Machinery.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler et al., 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.

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

Escudé Font, J., Costa-jussa, M., and R. (2019). Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy, August. Association for Computational Linguistics.

Field, A., Blodgett, S. L., Waseem, Z., and Tsvetkov, Y. (2021). A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online, August. Association for Computational Linguistics.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hanna, A., Denton, E., Smart, A., and Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. FAT* '20, page 501–512, New York, NY, USA. Association for Computing Machinery.

Haokip, T. (2021). From 'chinky' to 'coronavirus': racism against northeast indians during the covid-19 pandemic. *Asian Ethnicity*, 22(2):353–373.

Hasanuzzaman, M., Dias, G., and Way, A. (2017). Demographic word embeddings for racism detection on Twitter. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 926–936, Taipei, Taiwan, November.

Kaminsky; Long, R. D. (2011). *India Today: An Encyclopedia of Life in the Republic.* ABC-CLIO. ISBN 978-0-313-37463-0.

Manzini, T., Yao Chong, L., Black, A. W., and Tsvetkov, Y. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. 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 615–621, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Pinto, A. (2001). Un conference against racism: Is caste race? *Economic and Political Weekly*, 36(30):2817–2820.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*,

pages 15–20. Association for Computational Linguistics, June.

Zhou, P., Shi, W., Zhao, J., Huang, K.-H., Chen, M., Cotterell, R., and Chang, K.-W. (2019). Examining gender bias in languages with grammatical gender. 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)*, pages 5276–5284, Hong Kong, China, November. Association for Computational Linguistics.