Disentangling Societal Inequality from Model Biases: Gender Inequality in Divorce Court Proceedings

Sujan Dutta
Rochester Institute of Technology
sd2516@rit.edu

Parth Srivastava
Indian Institute of Technology, Kanpur
parthsri@iitk.ac.in

Vaishnavi Solunke
Rochester Institute of Technology
vs5709@rit.edu

Swaprava Nath
Indian Institute of Technology, Bombay
swaprava@cse.iitb.ac.in

Ashiqur R. KhudaBukhsh [*]
Rochester Institute of Technology
axkvse@rit.edu

Abstract

Divorce is the legal dissolution of a marriage by a court. Since this is usually an unpleasant outcome of a marital union, each party may have reasons to call the decision to quit which is generally documented in detail in the court proceedings. Via a substantial corpus of 17,306 court proceedings, this paper investigates gender inequality through the lens of divorce court proceedings. While emerging data sources (e.g., public court records) on sensitive societal issues hold promise in aiding social science research, biases present in cutting-edge natural language processing (NLP) methods may interfere with or affect such studies. We thus require a thorough analysis of potential gaps and limitations present in extant NLP resources. In this paper, on the methodological side, we demonstrate that existing NLP resources required several non-trivial modifications to quantify societal inequalities. On the substantive side, we find that while a large number of court cases perhaps suggest changing norms in India where women are increasingly challenging patriarchy, AI-powered analyses of these court proceedings indicate striking gender inequality with women often subjected to domestic violence.

Keywords: Gender Bias; LLM; Inconsistency Sampling

1 Introduction

The 2011 decennial census in India gave its citizens the following choices to select their marital status – never married, separated, divorced, widowed, married. Based on the census data, a study reported some startling facts [1]: 1.36 million of the Indian population is divorced which accounts for 0.24% of the married population, and 0.11% of the total population. More women were separated or divorced than men, and the number of separation was almost three times as high as the number of divorce.

Divorce, a historically taboo topic in India for ages [2], seldom features in mainstream Indian discourse [3]. Recent indications of changing social acceptance of divorcees notwithstanding [4], divorce in India still carries a considerable social stigma [5].

How do we quantify gender inequality in Indian divorce? Surveys about divorce often have limited participation and a small sample size [6], perhaps due to the social stigma attached. A vulnerable community – Indian women under conjugal distress – had limited visibility to social scientists. Via a substantial corpus of 17,306 divorce court proceedings, this paper conducts the first-ever computational analysis of gender inequality in Indian divorce based on public court records.

[*] This work is accepted at IJCAI 2023 (AI for good track). Ashiqur R. KhudaBukhsh is the corresponding author.

husband → wife	*wife → husband*
In the petition, it is alleged that the respondent always mentally and physically harassed the petitioner and threatened the petitioner that she will commit suicide and sometimes, she even physically tortured the petitioner and she used to beat and slap the petitioner when she becomes angry.	The respondent is a drunkard and he used to consume alcohol everyday and he ill treated and tortured the petitioner, demanding more dowry.
It was held that a husband cannot be expected to continue to live with the wife in the face of her sustained attitude of causing humiliation and calculated torture.	The interaction of the petitioner with her friends and relatives were viewed with suspicion. The petitioner’s telephone facility was disconnected. The Email account of the petitioner was checked by the respondent and often her friends and colleagues were threatened and abused.
According to him, he was working abroad till 2019 and only when he returned to India did he realise that he was cheated.	After considering the evidence, it was found that she was not willing to go with her husband further as he had cheated her once.

Table 1: *Snippets from court proceedings where unpleasant verbs (e.g., cheat, torture, slap, abuse, beat, threaten etc.) are used. The left column (husband → wife) accuses the wife of the wrongdoing. The right column (wife → husband) accuses the husband of the wrongdoing.*

Even though written in English, legal texts are often domain-specific [7]. The considerable variation of legal jargon across countries and courts makes domain-specific analysis important. In that vein, Indian legal NLP is an emerging field [7, 8]. Most NLP research on legal texts thus far has focused on building robust tools to analyze legal text. Recent research, however, on in-group bias [9] and sexual harassment [10], and Figure 1 and Table 1 suggest that automated methods to glean social insights from large-scale, legal texts merit investigation. Barring few recent lines of work [11, 12, 13], there is surprisingly little literature on large-scale linguistic analysis of gender bias in India, let alone on legal text zeroing in on divorce.

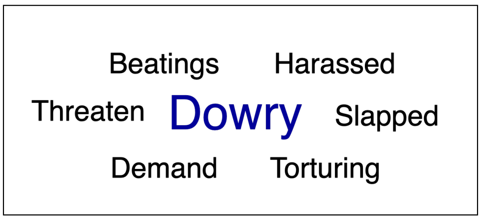


Figure 1: *Nearest neighbors of the word* `dowry` *in a word embedding trained on divorce court proceedings from Rajasthan, a state from India. Despite legal prohibition since 1961 [14], this retrograde social practice has continued in India with several studies linking it to other social crises such as female feticide, domestic abuse and violence, and dowry deaths [15].*

While emerging data sources (e.g., public court records available on the web) offer opportunities for social scientists to study important and sensitive social issues that previously had limited survey data, applying cutting-edge NLP methods to newer domains requires careful evaluation of the critical question: *How much of the (perceived) gender inequality as quantified by the methods truly reflects the corpus and how much of it is due to the inherent biases of the employed NLP methods?* In this paper, we show that the subtleties present in legal text present unique challenges. Unless we consider them and make non-trivial changes to existing methods, we may end up drawing inaccurate social conclusions. We further show that sophisticated NLP methods built on top of large language models (LLMs) need scrutiny when applied to social inference tasks involving genders. We, in fact, conduct a much broader *bias audit* of these systems. Our audit reveals well-known LLMs often exhibit gender bias even on simple subject-verb-object sentence completion tasks. Through a corpus-specific text entailment analysis, we demonstrate that downstream applications such as natural language inference (NLI) systems also exhibit sensitivity to gender. We finally, present a novel inconsistency sampling method to mitigate this bias and present our social findings.

To summarize, our contributions are the following:

Social: We create a substantial corpus of 17,306 divorce court proceedings and conduct the first-ever analysis of gender inequality through the lens of divorce proceedings. While a large number of court cases perhaps suggest changing norms in India where women are increasingly challenging patriarchy [16], our analyses reveal widespread domestic violence, dowry demands, and torture of the bride.

Methodological: We address extant gaps and limitations in multiple NLP frameworks. We propose non-trivial modifications to the WEAT framework [17] to make it suitable for legal text. We demonstrate a novel application of text entailment [18] in quantifying gender inequality. We investigate several potential sources for model bias in NLP resources that can interfere with quantifying gender inequality. We present a novel inconsistency sampling method exploiting counterfactuals to mitigate this bias.

2 Dataset

2.1 Collection

We scrape all the publicly available court proceedings with the word `divorce` between January 1, 2012 to December 31, 2021 from https://indiankanoon.org/ (hereafter IndK), an Indian law search engine launched in 2008 and the largest free online repository of the court proceedings of different courts and tribunals of India. Prior computational law research [19] and gender focused social science studies [10] have used IndK as source of data.

We download 86,911 case proceedings containing the word `divorce` from IndK using its advanced search feature. Filtering based on the keyword `divorce` is a high-recall approach to obtain relevant cases with precedence in computational social science research [20, 21]. However, the presence of the keyword `divorce` may not always indicate a divorce court proceeding; for instance, the keyword can be used to describe the marital status of any of the litigants. It can also be used in an altogether different context (e.g., *divorced from reality*). We use the following heuristic to further refine our dataset. We also look for other words (e.g., `husband`, `wife`, `marriage`) and phrases (e.g., `decree of divorce`), and check if such occurrences repeat for a minimum threshold (set to 5). On a random sample of 100 cases after we apply this cleaning method, a manual inspection reveals that 94 are divorce cases. Hence, our keyword-based filtering is reasonably precise. This pruning step retains 25,635 cases.

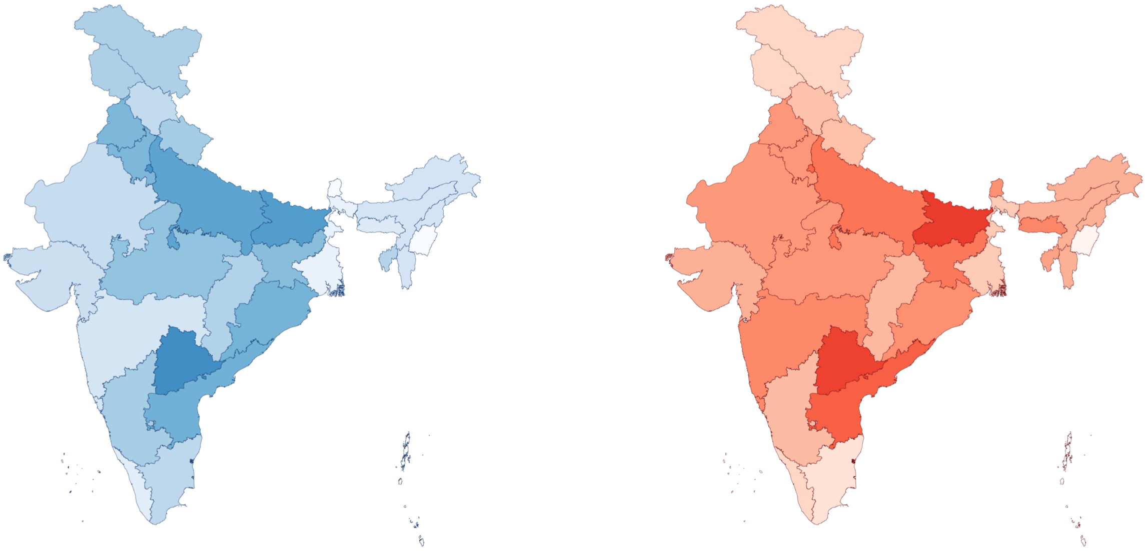
2.2 Data Pre-processing

To quantify gender inequality in court proceedings, we must disambiguate the legal parties – the plaintiff and the defendant – and accurately identify of the husband and the wife, who plays which role. Indian legal documents use a wide range of legal terms to denote the plaintiff (e.g., appellant, applicant, complainant, petitioner) and the defendant (e.g., respondent, nonapplicant, opponent). We observe different courts have different formats (sometimes, multiple formats) to summarize the proceedings. The documents also specify which party in marriage represents which role in several different ways (e.g., respondent/wife, respondent-wife, respondent aggrieved wife). We write a regular-expression-based pipeline and consolidate such information to identify the gender of the plaintiff and the defendant across all the states.

The names and salutations (e.g., `Mr.`, `Mrs.`, `Smt.`, `Shri`) of the plaintiff and defendant also provide gender information. Subcultural naming conventions played a key role in assigning gender to the litigants in some of the cases. For instance, `Kaur`, meaning princess, is a Punjabi last name only for females [22]. Or `ben`, meaning sister, is solely used in many female names in Gujarat [23]. Dependence information of the litigants also provides gender information (e.g., `son of`, `daughter of`, `wife of`).[1]

Of the 25,635 cases, we could unambiguously assign gender to 17,306 cases. For each case, we replace each mention of the litigants as `wife` or `husband` accordingly. For example, a proceeding snippet "*The plaintiff/wife has filed for a divorce. The plaintiff was married to the defendant for three*

[1] We did not find a single mention of `husband of` in our dataset.



(a) *Dowry*

(b) *IPC 498-A*

Figure 2: *Choropleths of divorce cases mentioning* `dowry` *and* `IPC 498-A`*. For each state, we compute the total number of divorce cases that mention* `dowry` *(for Figure 2a) or* `498-A` *(for Figure 2b) at least once and divide by the total number of divorce cases in that state. Each number is in the range [0, 1]. A larger number indicates greater mention of* `dowry` *or* `498-A`*. Higher intensity colors indicate larger values. Section 498-A is a section in the Indian Penal Code (IPC) introduced in 1983 to protect women from marital cruelty. The base maps used for this plot are sourced from the Government of India. The authors are aware that these maps include disputed territories. These maps do not constitute judgments on existing disputes.*

years.", will be modified to "*The wife has filed for a divorce. The wife was married to the husband for three years.*" This data set, $\mathcal{D}_{divorce}$, consists of 30,615,754 (30 million) tokens.

3 Brief overview of Indian legal system

Indian Judicial System is largely based on the English Common Law system (where, the law is developed by judges through their decisions, orders, and judgments). The nation has 28 states and 8 union territories (UT), and a total of 25 high courts (some high courts have jurisdiction of more than a state or UT). The federal structure has a supreme court coupled with the high courts that roughly handle the cases in a state or UT. The legal cases of divorce are usually handled by the family or district courts. However, some unresolved cases or sometimes fresh cases are also heard by the high courts. Since the court proceedings are public records and are digitally made available freely by IndK, we found this dataset to be quite appropriate for a large-scale study on gender equality in court proceedings.

4 Dowry in Divorce Proceedings

The dowry system involves a transaction of financial assets between the bride's family and the bridegroom's family with the latter being the recipient of the financial assets. While legally prohibited in India since 1961 [14], this practice has continued well after its legal prohibition and has a strong link to social crises such as female feticide [24], domestic abuse and violence [25, 26], and dowry deaths [15]. In order to protect the bride from marital cruelty and domestic violence, Indian Penal Code introduced Section 498 in 1983 [27].

Figure 2 reflects relative proportions of divorce cases containing the text tokens `dowry` and `498-A`. For each state, we report the fraction of divorce cases that contain at least one mention of these two tokens. A higher intensity color indicates a larger proportion of such cases. We observe that overall, 24.38% of all cases and 21.86% of all cases mention `dowry` and `498-A`, respectively. Jacob and Chattopadhyay, [1] reported that divorce in India does not follow any one-size-fits-all pattern across different states; there exists sufficient interstate variation even for the rate of divorce. We notice a considerable variation in mentions of dowry and section 498-A across different states indicating variance in reported cases of dowry or domestic violence. Among the states and the union territories, the top three entries in terms of dowry mentions are Telangana, Delhi, and Bihar while the top three entries in terms of Section 498-A mentions are Bihar, Telangana, and Andhra Pradesh. Bihar and Telangana have social science literature documenting dowry and domestic violence [28, 29]. Apart from the overlap in the top three entries, the statewise dowry and 498-A mentions are moderately correlated (correlation coefficient: 0.67).

We next conduct a qualitative analysis of (alleged) dowry demands [2]. On a random sample of 100 court proceedings where the (alleged) dowry demand is explicitly recorded, we observe that the estimated demanded amount is ₹393,100 ± 544,876. We observe demanded amounts as low as ₹5,000 to as high as ₹3,000,000 which explains the staggeringly high variance in our estimation. This also indicates the broad economic spectrum present in India and how far and wide the system of dowry (allegedly) persists. We further observe that cash is not always the solely demanded financial asset. Gold is the second-most commonly demanded asset. Out of the 100 cases, 34 cases report gold demands (71.2 ± 84.6 gm). When we adjust the valuation of demanded gold replacing it with the historical average gold price in India across 2012 and 2021 [3], the estimated (alleged) demanded dowry is ₹474,798 ± 567,219.

5 Methods Overview

We use two NLP methods to quantify gender inequality: (1) Word Embedding Association Test; and (2) a text entailment framework. A brief description follows.

5.1 Word Embedding Based Methods

The first metric is called **W**ord **E**mbedding **A**ssociation **T**est (WEAT) introduced by [17]. To calculate the metric, the words are embedded and the vectors $\vec{a}$ and $\vec{b}$ are obtained for the words a and b respectively. The cosine similarity of these words are denoted by $\cos(\vec{a}, \vec{b})$. The metric considers two sets of *target words* given by $\mathcal{X}$ and $\mathcal{Y}$, and two sets of *attribute words* $\mathcal{A}$ and $\mathcal{B}$. Then, the WEAT score is defined as $\text{WEAT}(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}) = (\text{mean}_{x \in \mathcal{X}} \sigma(x, \mathcal{A}, \mathcal{B}) - \text{mean}_{y \in \mathcal{Y}} \sigma(y, \mathcal{A}, \mathcal{B})) / \text{stddev}_{w \in \mathcal{X} \cup \mathcal{Y}} \sigma(w, \mathcal{A}, \mathcal{B})$, where, $\sigma(w, \mathcal{A}, \mathcal{B}) = \text{mean}_{a \in \mathcal{A}} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in \mathcal{B}} \cos(\vec{w}, \vec{b})$. Intuitively, $\sigma(w, \mathcal{A}, \mathcal{B})$ measures the association of w with the attribute sets, and the WEAT score measures the differential association of the two sets of target words with the attribute sets. A positive WEAT score implies that the target words in $\mathcal{X}$ is more associated with the attribute words in $\mathcal{A}$ than $\mathcal{B}$ and the words in $\mathcal{Y}$ is more associated with $\mathcal{B}$ than $\mathcal{A}$.

5.2 Text Entailment Based Methods

Quantifying gender inequality relying on the distributed representation of words presents a diffused, bird's-eye view of the larger trends. Also, these methods are known to be data-hungry [30]. Data availability often becomes a limiting factor to conducting contrastive studies at different spatio-temporal granularity. In what follows, we present a novel application of text entailment, a natural language inference (NLI) task [31] that bypasses the data size requirement and equips us with a finer lens through which we can compare and contrast gender inequality with respect to individual verbs.

An NLI system take a premise $\mathcal{P}$ and a hypothesis $\mathcal{H}$ as input and outputs entailment, contradiction, or semantic irrelevance. For instance, the hypothesis *some men are playing a sport* is entailed by the premise *a soccer game with multiple males playing* [32]. As one can see, textual entailment is more relaxed than pure logical entailment and it can be viewed as a human reading $\mathcal{P}$ would infer most likely $\mathcal{H}$ is true. This framework has gained traction in several recent social inference tasks that include estimating media stance on policing [33, 21], aggregating social media opinion on election fairness [34], and detecting COVID-19 misinformation [35]. Formally, let $NLI(\mathcal{P}, \mathcal{H})$ takes a premise $\mathcal{P}$ and a hypothesis $\mathcal{H}$ as inputs and outputs $o \in \{entailment, contradiction, neutral\}$. Following [21], we define entailment ratio (denoted by $ent(\mathcal{D}, \mathcal{H})$) for given corpus $\mathcal{D}$ and a hypothesis $\mathcal{H}$, as the fraction of the individual sentences present in $\mathcal{D}$ that entails $\mathcal{H}$: $ent(\mathcal{D}, \mathcal{H}) = \frac{\sum_{\mathcal{P} \in \mathcal{D}} \mathbb{I}(NLI(\mathcal{P}, \mathcal{H}) = entailment)}{|\mathcal{D}|}$, where $\mathbb{I}$ is the indicator function. A larger value of $ent(\mathcal{D}, \mathcal{H})$ indicates greater support for $\mathcal{H}$ in the corpus.

[2] This analysis follows the statements made by the plaintiffs

[3] Obtained from https://www.bankbazaar.com/gold-rate/gold-rate-trend-in-india.html

Consider we are interested in learning how often the husband and the wife are accused of torture (physical or emotional) in our corpus. We analyze this research question in the following way. We first construct a sub-corpus $\mathcal{D}_{torture}$ from the divorce court proceedings consisting of sentences that (1) mention `husband` or `wife` at least once; and (2) mention `torture` as a verb at least once. We next construct two hypotheses – $\mathcal{H}_{\texttt{MV},torture}$ and $\mathcal{H}_{\texttt{FV},torture}$ – using a `man` and a `woman` as victims and perpetrators interchangeably. $\mathcal{H}_{\texttt{MV},torture}$ is *A woman tortures a man* and $\mathcal{H}_{\texttt{FV},torture}$ is *A man tortures a woman*. We next compute the entailment gap defined as

$gap(\mathcal{D}_{torture}, torture) =$
$ent(\mathcal{D}_{torture}, \mathcal{H}_{\texttt{FV},torture}) - ent(\mathcal{D}_{torture}, \mathcal{H}_{\texttt{MV},torture})$

Effectively, this means we compute the fraction of sentences that entail *A woman tortures a man* in $\mathcal{D}_{torture}$ and subtract it from the fraction of sentences that entail *A man tortures a woman* in $\mathcal{D}_{torture}$. An overall positive number indicates that the male has been described as the torturer more often than the female in court proceedings. A negative value would indicate the opposite way. Similar analysis can be extended to other verbs such as `assault`, `beat`, or `abuse`.

6 Design Considerations

Adapting the `WEAT` and entailment frameworks to quantify gender inequality in our domain requires careful consideration of several aspects described in what follows.

6.1 Verbs for Target Sets

Traditionally, WEAT score is used to quantify gender or racial stereotypes. Majority of the elements present in those attribute sets would be nouns and adjectives (e.g., criminals, terrorists, doctors, police) [36, 37] and seldom verbs [38]. We are interested in understanding the action space of the two parties fighting a divorce case; we want to know if the court described that one party tortured or abused the other. Hence, verbs are a natural choice for our target set.

We inspect the list of high-frequency verbs in the corpus and narrow down to the following ten verbs: $\mathcal{X}_{unpleasant}$ = {`abuse`, `assault`, `beat`, `burn`, `cheat`, `misbehave`, `rape`, `slap`, `threaten`, `torture`}. A small subset of these words are already present in the list of unpleasant stimuli presented in [37]. We further compute the average valence score of these words as per the lexicon presented in [39]. We find the average valence score of $\mathcal{X}_{unpleasant}$ is 2.7, comparable to the average valence score (2.16) of unpleasant stimuli presented in [37].

Divorce being a bitterly fought family situation, we observe a sparse presence of pleasant verbs such as `love`, `care`, or `empathy` in our corpus. Since infrequent words in the corpus do not have reliable embeddings [40], in contrast with traditional applications of WEAT score, we choose the target set $\mathcal{Y}$ to be an empty set.

6.2 The Torturer and the Tortured

The attribute sets $\mathcal{A}$ and $\mathcal{B}$ as defined in the WEAT score represents the identifiers used for the plaintiff and defendant in our data (e.g., $\mathcal{A}$ consisting of `he`, `him`, `husband`, and $\mathcal{B}$ consisting of `she`, `her`, `wife` etc.). However, notice that WEAT score is agnostic about whether the identifier is the *contributor* or the *receptor* of target words. For example, torture does not happen in isolation; it requires a torturer and one who is tortured. Unlike nouns, verbs are typically associated with two entities – the subject and the object. To disambiguate between "*the husband tortured the wife*" and "*the wife tortured the husband*", a word embedding needs to understand this nuance. Otherwise, the embedding is likely to place both the plaintiff and defendant identifiers equidistant to the verb.

To disambiguate these two situations, we run the corpus through the `stanza` POS tagger [41] to find out the subject and object of the sentences and whether the statements are in active or passive voice. Based on this, we classify the subjects and objects as 'male perpetrator', 'female perpetrator', 'male victim', or 'female victim', in the sentences that has the target verbs. We replace these four cases with four unique words (denoted by w_{MP}, w_{FP}, w_{MV}, and w_{FV}, respectively) so that those words do not occur anywhere else in any of the documents. We call this new dataset $\mathcal{D}_{replaced}$.

7 Word Embedding Based Analysis

We are interested in two research questions:

***RQ 1:** How does gender inequality manifest in divorce court proceedings with respect to unpleasant verbs in $\mathcal{X}$?*

***RQ 2:** Is our careful disambiguation of the torturer and the tortured necessary at all?*

In order to answer these two questions, we run two sets of experiments with identical training configurations. First, we run experiments on $\mathcal{D}_{replaced}$ using the target and attribute sets as defined in the previous section. We train the word embedding model 10 times and calculate the WEAT scores for each of the following two cases: when both genders are (a) perpetrators, i.e., when $\mathcal{A} = \{w_{\text{MP}}\}, \mathcal{B} = \{w_{\text{FP}}\}$, and (b) victims, i.e., when $\mathcal{A} = \{w_{\text{MV}}\}, \mathcal{B} = \{w_{\text{FV}}\}$. We use the default parameters for training our `FastText` [42] Skip-gram embedding with the dimension set to 100 for all word-embeddings in this paper. Second, we run a *baseline* experiment with the original text data without replacing them with the four unique words ($\mathcal{D}_{divorce}$) and use the attribute sets as $\mathcal{A} = \{\texttt{husband}\}$ and $\mathcal{B} = \{\texttt{wife}\}$. The number of runs and the embedding method are the same in both experiments. The results are shown in Figure 3.

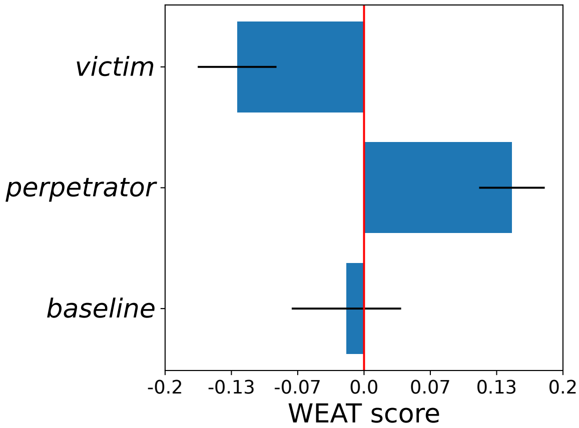


Figure 3: WEAT *scores. The* WEAT *score is averaged over ten runs. A larger positive value indicates a greater bias toward men. The top two values (victim and perpetrator) are computed on $\mathcal{D}_{replaced}$. The bottom (baseline) value is computed on $\mathcal{D}_{divorce}$. The top value (victim) is from the perspective of the victim where the attribute sets $\mathcal{A}$ and $\mathcal{B}$ are set to words denoting male victims and female victims, respectively. A negative value implies that women are more associated with the unpleasant verbs in $\mathcal{X}_{unpleasant}$. The middle value (perpetrator) is from the perspective of the perpetrator. A positive value implies that men are more associated with the unpleasant verbs in $\mathcal{X}_{unpleasant}$. The baseline indicates that without incorporating this nuance, the* WEAT *framework will present an inaccurate evaluation of the social bias present in the corpus.*

As already described, a negative WEAT score indicates $\mathcal{B}$ is more associated with the target set as compared to $\mathcal{A}$. Hence, if we look from the perspective of the victim, we find that women are more associated with the unpleasant verbs than men. In contrast, when viewed from the perpetrator's perspective, a positive WEAT score implies that men are more associated with the unpleasant verbs. Hence, our results indicate that in our corpus, women are more often the victims while men are more often the perpetrators.

Our baseline experiments that do not make any distinction between the perpetrator and the victim give a WEAT score close to zero indicating near-perfect gender equality. This inaccurate result, while highly surprising from a social science perspective, is not unexpected given how the original framework functions. The two entities (husband and wife) are present around the unpleasant verbs with nearly equal frequency. If the method does not make any distinction between the roles of victim and perpetrator, WEAT will give inaccurate results. We thus carefully use the WEAT score to elicit the correct gender bias when applied to legal texts for our social science research question.

8 Societal Inequality and Model Bias

Our word embeddings are computed from scratch while our next set of experiments relies on downstream applications built on top of large language models. Large language models (LLMs) are known to have a wide range of biases due to the train data [43] and extant literature has examined gender bias in the form of occupational stereotypes present in NLI systems [44]. We thus need to disentangle societal inequalities that are potentially reflected in our corpus and model biases that are potentially present in the NLP applications.

Essentially, for a premise/hypothesis pair $\langle \mathcal{P}, \mathcal{H} \rangle$, the NLI system estimates the probability P($\mathcal{H}$ $|\mathcal{P}$). However, how LLMs encode the probability P($\mathcal{H}$) when the hypotheses primarily consists of the two genders (male and female) and a set of verbs is understudied. A thorough investigation first reveals that the masked word prediction probability of several well-known LLMs is sensitive to gender. We next present a measure to quantify gender bias sensitivity of NLI frameworks and present mitigating strategies. Finally, we use a bias-mitigated NLI system on our corpus and report findings.

8.1 Implicit Bias in Agent and Theme in LLMs

Unlike existing literature that primarily target occupational stereotypes to quantify and analyze gender bias [44, 45, 13, 17, 46], we focus on a very basic unit in a sentence – the verbs. Following [47], let in a sentence *X verbs Y*, *X* represent the agent and *Y* represent the theme. Many verbs imply the relative authority levels between the agent and the theme. For example, in the sentence *The football coach instructed the players to play a conservative game*, the agent (the football coach) has more authority than the theme (the players). In contrast, the agent has less authority than the theme in the sentence *The football coach honored the players' suggestion to play a conservative game.* First proposed in [47], the connotation relation of power captures this notion of power differential between an agent and a theme with respect to a given verb.

While the connotation relation of power has been analyzed in the context of gender inequality in movie scripts [47] and follow-on research focused on editorial fixes to remove bias [48], little or no literature exists that documents the implicit gender bias present towards the agent and the theme when specific verbs are considered. This research is important and has a broader impact beyond our current social inference task. For instance, if an LLM encodes that it is less likely for a woman to inspire or guide someone than a man, this bias may percolate to downstream tasks leading to erroneous social conclusions when applied to large-scale data for other social inference tasks.

We use cloze tests to evaluate this implicit bias. A brief description of cloze test follows.

Cloze test: When presented with a sentence (or a sentence stem) with a missing word, a cloze task [49] is essentially a fill-in-the-blank task. For instance, in the following cloze task: *In the* `[MASK]`*, it snows a lot*, `winter` is a likely completion for the missing word. Word prediction as a test of LLM's language understanding has been explored in [50, 51].

Bias Evaluation Framework: We describe our proposed testing framework for gender bias. Let $\texttt{LLM}_{cloze}(w, \mathcal{S})$ denote the completion probability of the word w with a masked cloze task $\mathcal{S}$ as input. For a given verb v, we consider the following four cloze tests:

1. A [MASK] v a woman (denoted by $v_{womanAsTheme}$)
2. A [MASK] v a man (denoted by $v_{manAsTheme}$)
3. A man v a [MASK] (denoted by $v_{manAsAgent}$)
4. A woman v a [MASK] (denoted by $v_{womanAsAgent}$)

In an ideal world where the LLM treats men and women equally, $\texttt{LLM}_{cloze}(man, v_{womanAsTheme})$ and $\texttt{LLM}_{cloze}(woman, v_{manAsTheme})$ should be equal. However, our preliminary exploratory analysis indicates that is not the case. For example, when v is set to *inspire*, $\texttt{BERT}_{cloze}(man, v_{womanAsTheme})$ is 0.20 whereas $\texttt{BERT}_{cloze}(woman, v_{manAsTheme})$ is 0.16. When we set v to *guide*, the gap widens – $\texttt{BERT}_{cloze}(man, v_{womanAsTheme})$ is 0.71 whereas $\texttt{BERT}_{cloze}(woman, v_{manAsTheme})$ is 0.36.

Again, in an ideal world where the LLM treats men and women equally, $\texttt{LLM}_{cloze}(man, v_{womanAsAgent})$ and $\texttt{LLM}_{cloze}(woman, v_{manAsAgent})$ should be equal.

Let $\mathcal{V}$ denote the set of all verbs listed in [47] where the agent has more power than the theme. Our overall measures of implicit bias are: (a) $(1/|\mathcal{V}|) \cdot (\sum_{v \in \mathcal{V}}(\texttt{LLM}_{cloze}(man, v_{womanAsTheme}) - \texttt{LLM}_{cloze}(woman, v_{manAsTheme})))$, and (b) $(1/|\mathcal{V}|) \cdot (\sum_{v \in \mathcal{V}}(\texttt{LLM}_{cloze}(man, v_{womanAsAgent}) - \texttt{LLM}_{cloze}(woman, v_{womanAsAgent})))$.

Measure (a) quantifies $bias_{agent}$. A positive value indicates that the LLM encodes a man being in the position of agent likelier than a woman on expectation. Measure (b) quantifies $bias_{theme}$. A positive value indicates that the LLM encodes a man being in the position of theme likelier than a woman on expectation. We investigate three well-known LLMs for this audit: `BERT` [52]; `RoBERTa` [53]; and `Megatron` [54]. We consider 1,222 verbs listed in [47]. We also consider verbs in $\mathcal{X}_{unpleasant}$ for this

LLM	$\mathcal{V}$, $bias_{agent}$	$\mathcal{V}$, $bias_{theme}$	$\mathcal{X}_{unpleasant}$, $bias_{agent}$	$\mathcal{X}_{unpleasant}$, $bias_{theme}$
`BERT` [52]	0.32	0.01	0.23	-0.04
`RoBERTa` [53]	0.33	0.04	0.49	0.12
`Megatron` [54]	0.30	0.03	0.34	0.08

Table 2: *Implicit bias in agent and theme in LLMs. $\mathcal{V}$ denotes the set of 1,222 verbs present in Sap et el. [47] where the agent is identified to have more power than the theme. $\mathcal{X}_{unpleasant}$ denotes the set of ten unpleasant verbs considered in our study.*

study.

Table 2 summarizes our gender bias audit of LLMs with respect to verbs implying more power to the agent than the theme. We first note that for both verb sets, $bias_{agent}$ is substantially larger than $bias_{theme}$. This result indicates that men are considerably more likely to be considered as the agent when women is the theme and the verb implies that the agent has greater power than the theme. We also note that the completions favor mildly men over women even for the theme, however, the values are closer to 0.

8.2 Implicit Bias in NLI Systems

We describe our approach to quantify model bias in our NLI framework specific to our task. Consider we modify the sub-corpus $\mathcal{D}_{torture}$ to $\mathcal{D}_{torture}^{flipped}$ where the gender identifiers in each premise sentence are flipped to the equivalent identifier of the opposite gender. For instance, the premise *The wife tortured the husband both mentally and physically* will be modified as *The husband tortured the wife both mentally and physically*. Flipping gendered words to test bias through counterfactuals in the context of coreference resolution has been previously explored in [55]. We argue that if a premise in $\mathcal{D}_{torture}$ entails *A man tortures a woman*, the flipped premise in $\mathcal{D}_{torture}^{flipped}$ should entail *A woman tortures a man* instead in a gender-neutral NLI system. Hence the entailment gap for `torture` computed on $\mathcal{D}_{torture}$ should be equal in magnitude and opposite in polarity as the entailment gap computed on $\mathcal{D}_{torture}^{flipped}$. The NLI system's ($\mathcal{M}$) overall bias score with respect to verbs present in $\mathcal{X}_{unpleasant}$ is thus computed as $NLI_{bias}(\mathcal{M}, \mathcal{X}_{unpleasant}) = \sum_{v \in \mathcal{X}_{unpleasant}} \frac{abs((gap(\mathcal{D}_v, v) + gap(\mathcal{D}_v^{flipped}, v))}{|\mathcal{X}_{unpleasant}|}$. In simple words, for each verb, we compute the entailment gap ($value_1$) for the relevant sub-corpus and the flipped sub-corpus ($value_2$). We subtract $value_2$ from $value_1$ and take the absolute value of the sum. The bias score is the average value of this sum across all verbs: a score close to 0 indicates that the NLI system has a minimal bias, whereas larger values indicate greater bias.

Our baseline is an off-the-shelf NLI system from Allen NLP trained using `RoBERTa` (denoted by $\mathcal{M}_{base}$). We find that $NLI_{bias}(\mathcal{M}_{base}, \mathcal{X}_{unpleasant})$ is 0.27 [4].

8.3 Bias Mitigation Via Inconsistency Sampling

Active Learning is a powerful and well-established form of supervised machine learning technique [56] characterized by the interaction between the learner, aka the classifier, and the teacher (oracle or annotator). Each interaction step consists of the learner requesting the teacher the label of an unlabeled instance sampled using a given sampling strategy and augmenting the data set with the newly acquired label. Next, the classifier is retrained on the augmented data set. This sequential label-requesting and re-training process continues until some halting condition is reached (e.g., exceeded annotation budget or the desired classifier performance). At this point, the algorithm outputs a classifier, and the objective for this classifier is to closely approximate the (unknown) target concept in the future. The key goal of active learning is to reach a strong performance at the cost of fewer labels.

Some of the well-known sampling methods include uncertainty sampling [56], certainty sampling [57], and density-based sampling [58]. Beyond a static strategy, more complex strategies such as

[4] We note that a bias-aware NLI variant from Allen NLP has a better starting point (bias score 0.20) than the base model. However, the bias-aware model exhibits slower convergence than the base model when we conduct our active learning steps as discussed in Section 7.3. With identical experimental setting, after iteration 3, the bias-aware model improves its bias score to 0.133.

Data	$NLI_{bias}(\mathcal{M}, \mathcal{X}_{unpleasant})$
$\mathcal{M}_{base}$	0.269
Iteration 1	0.177 ± 0.021
Iteration 2	0.110 ± 0.024
Iteration 3	0.103 ± 0.023

Table 3: *Bias evaluation. Each iteration performs one round of inconsistency sampling and adds 60 samples to the train set. Performance is reported on five runs with different random seeds.*

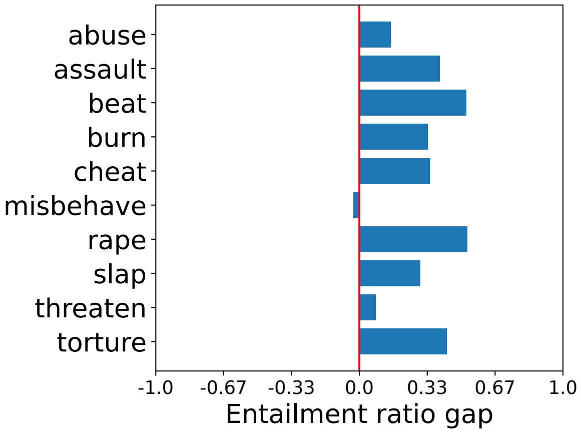


Figure 4: *Gender inequality using text entailment. For a given unpleasant verb, a negative value indicates that a female has played the role of a victim more often than a male.*

adapting strategy selection parameters based on estimated future residual error reduction or combining multiple sampling strategies to balance the label distribution in the procured data set have been explored in [59] and [60], respectively.

Inconsistency Sampling. First introduced in Dutta *et al.* [21], this sampling technique exploits the underlying logical structure of the $\langle premise, hypothesis \rangle$ space. For instance, a premise cannot both entail (or contradict) a given hypothesis and its negation. In our work, we extend this idea and exploit a $\langle premise, hypothesis \rangle$ space richer than Dutta *et al.* [21] for logical inconsistency.

Consider the premise/hypothesis pair *Continuously her husband used to harass and torture her everyday*/*A man tortures a woman*. We argue that if this premise entails the hypothesis (which it does), the modified premise/hypothesis pair with replacing every gendered word with the opposite gender – i.e., *Continuously his wife used to harass and torture him everyday*/*A woman tortures a man* – should also entail. If not, it signals a logical inconsistency. For each sampling iteration, we add 60 samples giving equal weightage to the verbs present in $\mathcal{X}_{unpleasant}$.

Table 3 summarizes our active learning results. For both models, $\mathcal{M}_{base}$ and $\mathcal{M}_{bias\text{-}aware}$, we conduct three rounds of active learning using inconsistency sampling and stop when the performance improvement becomes indiscernible ($\leq$ 0.01). All annotations are independently conducted by two annotators. Since legal documents are typically written in clear, unambiguous language, we observe a near-perfect agreement (Cohen's κ value 0.96). The remaining disagreements are resolved through a post-annotation adjudication step. Table 3 indicates that with subsequent active learning steps, our NLI system exhibits lesser bias. Given that the maximum possible bias score is 2, we achieve substantial improvement in mitigating the bias.

Now that we are more confident that our model inferences are less sensitive to gender, we evaluate the societal bias present in our corpus. Figure 4 summarizes our text entailment results. Barring `misbehave`, for all other verbs, men are identified as perpetrators more often than women. We further note that verbs that indicate physical abuse, such as `rape` and `beat`, particularly stand out with larger values. The average entailment gap for verbs unambiguously indicating physical harm – `assault`, `beat`, `burn`, `slap`, and `rape` – is much higher (0.41) than verbs that may or may not indicate physical harm (0.19) such as `abuse`, `cheat`, `misbehave`, `threaten`, and `torture`. A manual inspection of randomly sampled 200 $\langle premise, hypothesis \rangle$ pairs aligns with our automated method's overall findings.

9 Discussions and Limitations

In this paper, we present the first-ever computational analysis (to our knowledge) of gender inequality in divorce court proceedings in India. Based on the documented allegations of parties involved in the divorce, our analyses indicate a striking gender inequality as described in these public records. While documented evidence of marital distress in India exists in social science literature, how such factors play out in divorce has limited understanding. Our study sheds light on a vulnerable and vulnerable and practically invisible community in India.

Methodologically, we identify and address several gaps and limitations of existing NLP techniques to quantify gender inequality. We believe our finding specific to legal text is new, and our method to address it is simple, effective, and intuitive. Casting the problem of quantifying gender inequality as a text entailment task is also new. Our results on text entailment results suggest that NLI can be a viable tool to computational social science researchers to analyze similar research questions (e.g., who gets the child custody can be estimated with hypotheses *the husband gets the custody of the child* and *the wife gets the custody of the child*). Moreover, our bias mitigation strategy exploiting a novel inconsistency sampling technique using counterfactuals holds promise.

Our work has the following limitations.

Sentence level processing: An important point to keep in mind, however, is that our analyses operate at the sentence level. If in a court proceeding, a sentence records that the plaintiff accuses the defendant of wrongdoing which the defendant denies in a subsequent sentence, how these two contradicting claims are resolved in the court cannot be inferred without language models that can handle document-level contexts. We believe our research will open the gates for investigation with newer-age LLMs that can handle broader contexts.

Archival limitation: The sparse presence of the North-Eastern region in our dataset is most likely due to archival limitation as some of these states record the highest rate of divorce [1]. Our study is also limited by the overall archival extent of IndK.

Economic independence: Some of the court proceedings mention the litigants' occupations. We annotated randomly 100 sampled occupations for women. While an overwhelming majority of the sampled occupations are homemakers, compared to World Bank Data on labor force participation of women in India (23%), 32% of the women are working women in our sampled occupations. Economic independence and divorce merit a deeper exploration.

Out-of-court settlements, separation, abandonment: Finally, not all unhappy marriages end up in divorce and reach court for dissolution. Many out-of-court settlements happen. As documented in [1], the number of separated women in 2011 is almost three times the number of divorced women. Since divorce is still looked at as a social stigma [5] and family institutions are highly valued in India, there could be many women who continue with their dysfunctional marriages while unhappy. The court does not know their stories.

10 Ethical Statement

We work with public court records. Prior studies exist on Indian court proceedings [9]. We conduct aggregate analysis refraining from presenting any personally identifiable information in the paper. Hence, we do not see any ethical concern. Rather, we believe our findings and methods can be valuable to policymakers and social scientists.

A study on binary gender inequality runs the risk of oversimplifying gender, which we acknowledge lies on a spectrum. Same-sex marriage is yet not legal in India. Further nuances will be needed to extend our work to other cultures allowing same-sex marriages. We are also sensitive to previous studies that point out the potential harms of the erasure of gender and sexual minorities [61].

References

[1] Suraj Jacob and Sreeparna Chattopadhyay. Marriage dissolution in india: Evidence from census 2011. *Economic and Political Weekly*, 51(33):25–27, 2016. (Cited on pages 1, 4, and 11)

[2] Premchand Dommaraju. Divorce and separation in india. *Population and Development Review*, pages 195–223, 2016. (Cited on page 1)

[3] William J. Goode. Marital satisfaction and instability-a cross-cultural class analysis of divorce rates. *International social science journal*, 14(3):507–526, 1962. (Cited on page 1)

[4] A Santhosh Mani and Bhanu Priya. A study on the recent trends of divorce in india. *ZENITH International Journal of Multidisciplinary Research*, 7(8):25–32, 2017. (Cited on page 1)

[5] Jyothsna Belliappa. *Gender, class and reflexive modernity in India*. Springer, 2013. (Cited on pages 1 and 11)

[6] Bindhu Vasudevan, Devi M. Geetha, Anitha Bhaskar, Binu Areekal, Anupa Lucas, et al. Causes of divorce: a descriptive study from central kerala. *Journal of evolution of medical and dental sciences*, 4(20):3418–3427, 2015. (Cited on page 1)

[7] Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. A comparative study of summarization algorithms applied to legal case judgments. In *ECIR*, pages 413–428. Springer, 2019. (Cited on page 2)

[8] Arvind Kalia, Naveen Kumar, and Nischay Namdev. Classifying case facts and predicting legal decisions of the indian central information commission: a natural language processing approach. In *Advances in Deep Learning, Artificial Intelligence and Robotics*, pages 35–45. Springer, 2022. (Cited on page 2)

[9] Elliott Ash, Sam Asher, Aditi Bhowmick, Sandeep Bhupatiraju, Daniel Chen, Tanaya Devi, Christoph Goessmann, Paul Novosad, and Bilal Siddiqi. In-group bias in the Indian judiciary: Evidence from 5 million criminal cases. Technical report, Working paper, August, 2021. (Cited on pages 2 and 11)

[10] Anil Kumar. Sexual harassment of women at workplace: How far is indian law protective? *International Academic Journal of Law*, 1(1):35–39, 2020. (Cited on pages 2 and 3)

[11] Nishtha Madaan, Sameep Mehta, Taneea Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. Analyze, detect and remove gender stereotyping from bollywood movies. In *MAccT*, pages 92–105. PMLR, 2018. (Cited on page 2)

[12] Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, 2020. (Cited on page 2)

[13] Kunal Khadilkar, Ashiqur R. KhudaBukhsh, and Tom M. Mitchell. Gender bias, social bias, and representation in Bollywood and Hollywood. *Patterns*, 3(4):100486, 2022. (Cited on pages 2 and 8)

[14] R. Jaganmohan Rao. Dowry system in India — a socio-legal approach to the problem. *Journal of the Indian Law Institute*, 15(4):617–625, 1973. (Cited on pages 2 and 4)

[15] Nehaluddin Ahmad. Dowry deaths (bride burning) in India and abetment of suicide: a socio-legal appraisal. *JE Asia & Int'l L.*, 1:275, 2008. (Cited on pages 2 and 4)

[16] Reeta Sonawat. Understanding families in india: A reflection of societal changes. *Psicologia: Teoria e Pesquisa*, 17:177–186, 2001. (Cited on page 3)

[17] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. (Cited on pages 3, 5, and 8)

[18] Bill MacCartney and Christopher D. Manning. Modeling semantic containment and exclusion in natural language inference. In *COLING 2008*, pages 521–528, 2008. (Cited on page 3)

[19] Arpan Mandal, Kripabandhu Ghosh, Saptarshi Ghosh, and Sekhar Mandal. Unsupervised approaches for measuring textual similarity between legal court case reports. *Artificial Intelligence and Law*, 29(3):417–451, 2021. (Cited on page 3)

[20] Andrew Halterman, Katherine A. Keith, Sheikh Muhammad Sarwar, and Brendan O'Connor. Corpus-Level Evaluation for Event QA: The IndiaPoliceEvents Corpus Covering the 2002 Gujarat Violence. In *ACL/IJCNLP 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4240–4253, 2021. (Cited on page 3)

[21] Sujan Dutta, Beibei Li, Daniel S. Nagin, and Ashiqur R. KhudaBukhsh. A murder and protests, the capitol riot, and the chauvin trial: Estimating disparate news media stance. In *IJCAI*, pages 5059–5065, 2022. (Cited on pages 3, 5, and 10)

[22] Harjnder Kaur-Aulja, Farzana Shain, and Alison Lilley. A Gap Exposed: What is Known About Sikh Victims of Domestic Violence Abuse (DVA) and Their Mental Health? *European Journal of Mental Health*, 14(1):179–189, 2019. (Cited on page 3)

[23] PJ Mistry. Personal names: Their structure, variation, and grammar in Gujarati. *South Asian Review*, 6(3):174–190, 1982. (Cited on page 3)

[24] Devaki Monani Ghansham. Female foeticide and the dowry system in India. In *Townsville International Women's Conference, James Cook Univ., Australia*, 2002. (Cited on page 4)

[25] Priya R. Banerjee. Dowry in 21st-century India: the sociocultural face of exploitation. *Trauma, Violence, & Abuse*, 15(1):34–40, 2014. (Cited on page 4)

[26] Mudita Rastogi and Paul Therly. Dowry and its link to violence against women in India: Feminist psychological perspectives. *Trauma, Violence, & Abuse*, 7(1):66–77, 2006. (Cited on page 4)

[27] Deepshikha Carpenter and Polly Vauquline. Protecting Women from Domestic Violence in Assam, India? Evaluating Section 498-A, The Indian Penal Code (IPC), 1983 vs the Protection of Women from Domestic Violence Act (PWDVA), 2005. *Journal of International Women's Studies*, 18(1):133–144, 2016. (Cited on page 4)

[28] Gopalan Retheesh Babu and Bontha Veerraju Babu. Dowry deaths: a neglected public health issue in India. *International health*, 3(1):35–43, 2011. (Cited on page 5)

[29] Tanya Jakimow. 'everyone must give': Explaining the spread and persistence of bridegroom price among the poor in rural telangana, india. *Journal of Asian and African Studies*, 48(2):180–194, 2013. (Cited on page 5)

[30] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013. (Cited on page 5)

[31] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer, 2005. (Cited on page 5)

[32] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015. (Cited on page 5)

[33] Andrew Halterman, Katherine Keith, Sheikh Sarwar, and Brendan O'Connor. Corpus-level evaluation for event QA: The IndiaPoliceEvents corpus covering the 2002 Gujarat violence. In *ACL-IJCNLP*, pages 4240–4253, 2021. (Cited on page 5)

[34] Ashiqur R. KhudaBukhsh, Rupak Sarkar, Mark S. Kamlet, and Tom M. Mitchell. Fringe news networks: Dynamics of US news viewership following the 2020 presidential election. In *ACM WebScience*, pages 269–278, 2022. (Cited on page 5)

[35] Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, December 2020. (Cited on page 5)

[36] Thomas Manzini, Yao Chong Lim, Alan W. Black, and Yulia Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *NAACL-HLT*, pages 615–621, 2019. (Cited on page 6)

[37] Anthony G. Greenwald and Thomas F. Pettigrew. With malice toward none and charity for some: ingroup favoritism enables discrimination. *American Psychologist*, 69(7):669, 2014. (Cited on page 6)

[38] Alexander Hoyle, Lawrence Wolf-Sonkin, Hanna M. Wallach, Isabelle Augenstein, and Ryan Cotterell. Unsupervised discovery of gendered language through latent-variable modeling. In *ACL 2019*, pages 1706–1716, 2019. (Cited on page 6)

[39] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 2013. (Cited on page 6)

[40] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *ICLR*. OpenReview.net, 2018. (Cited on page 6)

[41] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *ACL: System Demonstrations*, 2020. (Cited on page 6)

[42] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 5:135–146, 2017. (Cited on page 7)

[43] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *ACM FaccT*, pages 610–623, 2021. (Cited on page 7)

[44] Rachel Rudinger, Chandler May, and Benjamin Van Durme. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, 2017. (Cited on page 8)

[45] Shanya Sharma, Manan Dey, and Koustuv Sinha. Evaluating gender bias in natural language inference. *CoRR*, abs/2105.05541, 2021. (Cited on page 8)

[46] Vaibhav Kumar, Tenzin Singhay Bhotia, and Tanmoy Chakraborty. Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. *TACL*, 8:486–503, 2020. (Cited on page 8)

[47] Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. Connotation frames of power and agency in modern films. In *EMNLP 2017*, pages 2329–2334, 2017. (Cited on pages 8 and 9)

[48] Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. Powertransformer: Unsupervised controllable revision for biased language correction. In *EMNLP 2020*, pages 7426–7441, 2020. (Cited on page 8)

[49] Wilson L. Taylor. "Cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953. (Cited on page 8)

[50] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *ACL 2016*, pages 1525–1534, 2016. (Cited on page 8)

[51] Allyson Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *TACL*, 8:34–48, 2020. (Cited on page 8)

[52] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. (Cited on pages 8 and 9)

[53] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. (Cited on pages 8 and 9)

[54] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *CoRR*, abs/1909.08053, 2019. (Cited on pages 8 and 9)

[55] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer, 2020. (Cited on page 9)

[56] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. (Cited on page 9)

[57] Vikas Sindhwani, Prem Melville, and Richard D. Lawrence. Uncertainty sampling and transductive experimental design for active dual supervision. In *ICML*, pages 953–960. ACM, 2009. (Cited on page 9)

[58] Hieu T. Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *ICML*, page 79, 2004. (Cited on page 9)

[59] Pinar Donmez, Jaime G Carbonell, and Paul N Bennett. Dual strategy active learning. In *Machine Learning: ECML 2007*, pages 116–127. Springer, 2007. (Cited on page 10)

[60] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. Voice for the voiceless: Active sampling to detect comments supporting the Rohingyas. In *AAAI 2020*, volume 34-01, pages 454–462, 2020. (Cited on page 10)

[61] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M. Phillips, and Kai-Wei Chang. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *EMNLP*, pages 1968–1994, 2021. (Cited on page 11)