Justice for the Disadvantaged: A Study of Public Reactions on Indian Supreme Court Judgments

Soumilya De[1]*, Soumyajit Datta[2], Koustav Rudra[3], Saptarshi Ghosh[3], Ashiqur KhudaBuksh[2], and Kripabandhu Ghosh[1]

[1] Indian Institute of Science Education and Research, Kolkata, India
*corresponding author: soumilya.de.scholar@gmail.com
[2] Rochester Institute of Technology, Rochester, New York, USA
[3] Indian Institute of Technology, Kharagpur, India

Abstract. The judgments of a country's apex court on socially pertinent issues often invite a wide spectrum of public reactions reflecting a range of polarities. In this paper, we examine public reactions to three landmark judgments of the Indian Supreme Court – *Triple Talaq* (on Islamic divorce), *Section 377* (focused on the criminalization of homosexual activity), and *Sabarimala Temple* (on restricting the entry of women and girls of reproductive age into a Hindu temple) – involving disadvantaged groups such as women and sexual minorities. To our knowledge, this is the first-ever work that investigates social web discourse pivoting landmark rulings for disadvantaged groups in India. We first annotate a substantial novel dataset of 23,418 comments, partially annotated through LLM-human partnership, particularly enriched by the participation of disadvantaged groups in the annotation process. Our analyses reveal that not all verdicts receive comparable support or criticism – civic receptivity considerably varies across verdicts, with the ban on instant Triple Talaq receiving overwhelming support.[4]

Keywords: disadvantaged groups, landmark verdicts, civic receptivity

1 Introduction

In 1829, Raja Ram Mohun Roy, a pioneering social reformer in India, submitted a petition to abolish '*Sati*' – a retrograde social practice of self-immolation of the widow on the funeral pyre of her husband [32]. Roy had a modest 300 signatories indicating their support in his petition. In contrast, the petition was challenged by a counter-petition that had more than 3,300 signatures [12]. Legal reforms for disadvantaged groups have seldom been smooth sailing and often faced resistance as they challenged religious beliefs and societal norms.[5]

[4] Note that the authors express *no* opinion on the verdicts of the honorable Supreme Court of India, and limit the study to analyzing *anonymous* public discourse. Some of the contents in this paper can be disturbing and deemed offensive.

[5] The Problems of Marginalized Groups in India; Living with Dignity: Sexual Orientation and Gender Identity-Based Human Rights Violations in India

Classes	Triple Talaq	Section 377	Sabarimala
Direct Critic	Such a shameless verdict ...	The greatest disaster. Shame on the judgment. Don't understand what these judges are sitting for??	"Notions of rationality cannot be invoked in matters of religion" SC should not interfere in religious matters
Indirect Critic	whatever they said everything is correct. don't interfere in religions matter.	Destruction of Indian society has started itself. Results shall be coming soon..	A temple is not a picnic spot
Indirect Favor	great relief to Muslim sisters as they were used, exposed, and disposed like tissue	*Finally, I am not a criminal by the law !* But an equal citizen of this great nation!	if a God punishes his followers just bcoz women want to worship him in his temple, he is not worthy to be a God
Direct Favor	Landmark judgement by honorable Supreme Court. Wish all Muslim women good luck! Justice for all!	This is the good decision taken by Supreme Court ... Bcoz Fundamental right are equal for all.	I always love to see the supremacy of the Constitution over all
Undetermined	When penguins defends sharks	Wow happy to know different opinions ,	it is situated in the Periyar Tiger Reserve

Table 1: Example comments per class. The first four classes capture stance relevant to the verdicts.

How does modern society react to watershed court judgments championing the rights of the disadvantaged? In this paper, via substantial datasets of 23,418 comments in total from 872 relevant YouTube videos, we analyze civic engagement with three recent major judgements of the Indian Supreme Court (INSC) favoring rights for disadvantaged groups: (1) abolishing the practice of instant Triple Talaq, (2) writing down of Section 377 focused on queer population, and (3) allowing entry of menstruating women into Sabarimala Temple (detailed in Section 1.1).

Several prior works have explored civic engagement as a stance detection problem on contemporary social issues ranging from gun control/right to the #metoo movement [16–19,29]. Notable works [24,31,34,35,39] in computational law in India revolve around legal documents. However, barring a few recent computational social science studies that examined gender and in-group biases in court proceedings [2, 14], civic reaction to important court rulings in India through the lens of the social web is rather underexplored. Table 1 indicates that social web discourse around controversial court rulings can present an effective instrument to gauge civic receptivity, especially if the decisions affect women and sexual minorities.

Our work introduces a novel aspect of stance detection where social media intersects with legal discourse. Unlike traditional stance detection datasets / approaches which have *three* classes (favor, against and neutral), we consider an additional dimension by distinguishing between direct and indirect references

to the judiciary for both favor and criticism, and consider *five* classes. People often express their opinions through allusions circumventing the use of definitive keywords related to the judiciary or the verdict, necessitating the granularity and adding to the complexity of the task.

Our Contributions: We make the following contributions in this paper:

- *Social:* Via a substantial corpus of 23,418 YouTube comments on 872 relevant YouTube videos, we analyze civic engagement with three major recent Supreme Court rulings in India. To our knowledge, this is the *first paper* that investigates social web discourse around watershed court rulings championing women and sexual minorities. Our analyses reveal that not all verdicts receive comparable support or criticism – civic receptivity considerably varies across verdicts, with the ban on triple talaq receiving overwhelming support.
- *Resource:* We release a novel dataset[6] of 23,418 social web posts (anonymized) with an annotated stance on the three verdicts. One of the annotators of our datasets self-identifies as a member of the queer community. Our work thus contributes to the growing literature of participatory AI [7,10,20,27,41], where disadvantaged stakeholders take prominent roles in the curation of the datasets and AI systems.
- *Methodological:* We leverage state-of-the-art NLP methods, customized for the novel task of *automatic* public stance detection toward Supreme Court verdicts. We address class imbalance by leveraging an LLM-assisted approach that augments new instances without compromising the organic nature of user-generated comments. Although our work considers three Indian Supreme Court judgments, the pipeline adopted can be deployed for other such rulings in India and other countries to gauge public reaction to controversial judgments.

1.1 Background

The Supreme Court of India has a history of delivering reformative judgments that drive social progress while balancing the sensitivities of a pluralistic society.

- ***Instant Triple Talaq:*** In 2016, Shayara Bano filed a writ petition in the Supreme Court of India when her marital status was altered in an instant from *married* to *divorced* through *Talaq-e-Biddat* or instantaneous Triple Talaq. A year later, the apex court described the pronouncement of divorce through successive utterances of Talaq as *manifestly arbitrary* and instituted a ban on it.[7] The 2017 verdict is seen as a notable attempt to grant justice to a historically oppressed gender practicing a specific religion [47] (approximately, 7% of the Indian population are Muslim women). In 2019, following the verdict, a bill was passed legislatively making the practice a criminal offense [4].
- ***Section 377:*** Pronounced under "Unnatural Offenses", Section 377 of the Indian Penal Code (instituted in 1860 during the British rule) stated that *whoever voluntarily has carnal intercourse against the order of nature with any man, woman or animal* may face *imprisonment for life*. It has been claimed that

[6] https://github.com/khorg0sh/Justice-for-the-Disadvantaged.
[7] Triple Talaq Explained

throughout its history, the section, with no explicit mention of sexuality, served as a tool for harassing and intimidating sexual minorities [33, 42].

In the absence of comprehensive data, and based on global prevalence estimates, the queer population of India is conservatively approximated to be at least 10% of the population, amounting to around 140 million. In a longed-for moment of respite for the huge queer population, the Indian Supreme Court in September 2018 ruled that consensual sexual activities among adults are no longer a criminal activity and partially read down Section 377 with a note *"History owes an apology to the members of this community"*. The sexual identities earlier associated with deviant behavior and offense now received recognition. The verdict has paved the way for further discussions on advancing LGBTQ+ rights in India – legalization of same-sex marriage,[8] right to adoption for LGBTQ+ couples,[9] and enactment of anti-discrimination laws.[10]

- ***Sabarimala Temple:*** Does an *exception placed on women because of biological differences violate the Constitution*? The Indian Supreme Court in the Sabarimala verdict passed in 2018 responded, stating that such exceptions are violative of the *Right to Equality* of women and perpetuate gender-based discrimination [3]. The temple of Lord Ayyappa at Sabarimala adhered to a practice believed to be a tradition [40] that prohibited entry of women of menstruating age (10-50). The justification for the ban is that Lord Ayyappa is said to be celibate (Britannica). In a 4:1 majority 2018 verdict, the apex court lifted the ban, citing that *the menstrual status of a woman cannot be a valid constitutional basis to deny her the dignity of being and the autonomy of personhood.* This verdict is seen as another attempt at ensuring gender equality within the purview of existing religious practices through legal alterations.

What binds these three verdicts? The *Right to Equality.* Article 14,[11] states that *The State shall not deny to any person equality before the law or the equal protection of the laws.* Across the cases, *discrimination* is a recurring theme where *equality* is a challenged prospect and the verdicts uphold the rights of the historically disadvantaged groups.

For the rest of this paper, we indicate these verdicts by italicizing them: *Triple Talaq* denotes the verdict that abolished instant Triple Talaq by law; *Sabarimala* denotes the verdict allowing entry of women into Sabarimala temple; and *Section 377* represents the verdict writing down Section 377.

2 Data Collection

To analyze the public discourse on the three aforementioned landmark judgments, we first construct the datasets from YouTube[12] comments on relevant

[8] Same Sex Marriages in India
[9] Why LGBTQIA+ couples should be allowed to adopt
[10] India's LGBTQIA+ community notches legal wins but still faces societal hurdles
[11] Indian Kanoon: Article 14
[12] www.youtube.com

Dataset	Total Comments	Comments in English
Triple Talaq	25,335	11,071
Section 377	9,141	5,350
Sabarimala	12,740	6,997

Table 2: The total number of comments collected vs the number of comments with English as the identified language

videos. To our knowledge, the datasets are the first of its kind with a novel set of classes. In what follows, we describe our dataset curation steps.

2.1 Data Collection

Our choice of social web platform (YouTube) is guided by (1) *YouTube's popularity in India*: As of 2024, YouTube has the highest user base in India, 462 million out of 820 million active internet users. (2) *Presence of relevant discourse*: Following a court ruling, various news agencies usually upload a variety of videos on YouTube covering the judgment. These videos, often presenting contrasting opinions, become focal points for discussion through comments. (3) *Ease of API access*: YouTube provides publicly available and free API access. *Each* dataset is prepared through the following steps :

1. **Keywords for search**: We curate keywords consisting of two components, the case name (e.g. *Sabarimala Temple*) and a supporting term from the set {*judgment*, *verdict*, *Supreme Court judgment*, *Supreme Court verdict*} trailing it. Therefore, per judgment we have four keywords to iterate over.
2. **API Search:** An API search is conducted on each keyword to **extract *relevant* (API parameter) videos**. The union of the video IDs is taken for further processing.
3. **Timeline-based filtration**: We observe a substantial number of videos are posted promptly following the verdict, but the initial enthusiasm gradually dies down. Hence, we consider only the videos posted in the first three months from the date of the verdict.
4. **Retrieval of comments**: The filtered video IDs are now processed to **retrieve comments** under each one of them, with associated details such as number of likes and replies. Note that we do *not* collect any user identities of the users who posted the comments.

Language of comments. The retrieved comments are in English and various Indian languages such as Hindi, Bangla, Malayalam, Telugu, etc. We consider only the lingua franca, i.e., English, mainly due to annotator unavailability for other Indian languages and potential region-specific bias in opinions in considering only select languages. To **identify and extract English comments**, we use Google Translator API version 4.0.0rc1. Table 2 reports the numbers.

2.2 Class definitions

On a subset of comments on individual verdicts, we first follow an *Open Coding* [25] approach and observe the following two broad comment patterns:

- **Direct:** References the judiciary *directly* using related terms such as 'Supreme Court', 'verdict', 'judgment', 'judges', etc. The keyword(s) used in the comment is the indicative factor. Identification of such comments requires no prior knowledge about the judgments.
- **Indirect:** Alludes to the judgment or the judiciary *indirectly* in no immediate judiciary-related term, and only suggests a stance. Such comments come with varied degrees of subjectivity where familiarity with the verdict and contemporary socio-political context is necessary (see examples in Table 1).

For instance, "shame on the judges" is a *Direct* comment where the categorization can immediately be based on the keyword "judges". In contrast, the comment "a temple is not a picnic spot" does *not* use any judiciary-related term and as per understanding of the verdicts, the comment refers to the Sabarimala temple, therefore, it is an example of *Indirect* comment.

Besides *Direct*/*Indirect*, the polarity is captured as:

- **Favor:** the stance expressed is of approval towards the verdict
- **Critic:** the stance expressed is of disapproval towards the verdict

Accordingly, we formulate the following classes based on the above-mentioned dimensions (see examples in Table 1):
1. Direct Critic: expresses criticism with direct reference to the judiciary.
2. Indirect Critic: criticism with *no* direct reference.
3. Direct favor: expresses approval with direct reference to the judiciary.
4. Indirect favor: approval with *no* direct reference.
5. Undetermined / Others
The first four classes reflect a clear stance regarding the verdicts while the fifth class *Undetermined/Others* considers samples where either the stance is unclear or there is no established relevance to the judgment. As an example, "Wow happy to know different opinions" bears no stance around the verdict, it merely expresses delight in observing a spectrum of opinions.

Candidates for Annotation: Consistent with prior literature around non-native English-speaking populations [36], we observe that a considerable portion of the collection exhibits spelling and grammatical disfluencies, consequently making the expression of the stance ambiguous. In our work, the expressiveness of a comment is assumed to be determined by public understanding of the comment that, in turn, is indicated by *engagement* in terms of likes and replies. Comments with a degree of engagement, as per our observation, are coherent and offer more clarity on the stance.

The engagement is quantified taking into account the number of likes and replies on each comment. We leverage the sum of the counts (of likes and replies) to arrange the comments in descending order. To obtain the set to annotate, we consider the 75 percentile value (Q3) as the threshold and retrieve only the comments with higher engagement (see Table 3 for the number of such comments), while the rest of the collection remains unlabeled.

Verdict	Q3 of L+R	# of comments
Sabarimala	1.0	1,655
Section 377	4.0	1,146
Triple Talaq	1.0	2,350
Total	-	5,151

Table 3: Number of candidate comments for annotation. L and R represent the number of likes and replies respectively. E.g., *Section 377* has 1,146 comments with L+R>4.0

Classes	Triple Talaq		Section 377		Sabarimala	
	Before	After	Before	After	Before	After
Direct Critic	12 (0.51%)	17 (0.68%)	30 (2.62%)	80 (5.01%)	107 (6.47%)	221 (11.81%)
Indirect Critic	95 (4.04%)	155 (6.24%)	193 (16.84%)	488 (20.58%)	466 (34.51%)	466 (24.91%)
Direct favor	84 (3.57%)	151 (6.08%)	130 (11.34%)	235 (14.72%)	42 (2.54%)	80 (4.28%)
Indirect favor	407 (17.31%)	407 (16.39%)	348 (30.37%)	348 (21.80%)	101 (7.45%)	165 (8.82%)
Undetermined	1752 (74.55%)	1752 (70.60%)	445 (38.83%)	445 (27.88%)	939 (56.74%)	939 (50.19%)

Table 4: The initial class distributions (Before) are enhanced to obtain comparatively less imbalanced datasets (After).

Annotation: The annotation process involved three annotators who are fluent in English and use of Youtube.[13] The initial annotation was done by two annotators where one annotator is a self-identified queer person, and the other identifies as a cisgender female. The senior annotator (the tie-breaker) identifies as cis-gender male. Prior literature has considered several approaches for disagreement resolution – e.g., majority voting [9, 48] or third objective instance [15]. In our work, accounting for the participation of the disadvantaged groups, we resolve the disagreements with a priority placed on the perspective of the annotator from the concerned disadvantaged group (e.g., our queer annotator for Section 377), in *Indirect* vs *Undetermined* where the majority of disagreement occurs. In case of a disagreement around the polarity, we use *third objective instance* and involve the third annotator. We observe considerable agreement in the annotation process (Cohen's κ 0.73).

2.3 Mitigating imbalance in the dataset

Table 4 shows that the class *Undetermined* consumes a significant share from each dataset while several classes with notably low numbers of comments lead to considerable imbalance in the datasets. To address this imbalance, we utilize an LLM-assisted approach that supplements the set of underrepresented classes with fresh user-generated comments, where the minimum requirement of *additional* annotator involvement is deferred to the final stage. While LLM-assisted annotation has been extensively studied as an alternative to human annotators [21, 26, 38], recent lines of work have also recommended caution [43]. Here,

[13] To be noted, given the offensive nature of several comments, the annotators were cautioned about potential sensitivity.

Background: factual information about the verdict.

Definition of Class `<class name>`: description of the class to be upscaled

Example Comments:
Example: `<example 1>`
Rationale: rationale why the example above belongs to the class `<class name>`

...

Candidate Comments:
1. `<candidate comment 1>`
2. `<candidate comment 2>`

...

Considering the background and the class description, which among the indexed candidate comments belong to the class `<class name>`. Return the indices of the chosen comments.

Fig. 1: The five-segment prompt contains a total of ten examples with rationales for an individual minority class and five candidates per comment in the minority class.

LLM is utilized to address *only* the imbalance in the datasets with the primary goal of minimizing human efforts.

We follow a two-stage approach for each dataset $\mathcal{D}$, where the first stage involves LLM and the second stage involves the human annotators.

The First Stage: LLM.

In the first stage, LLM receives through prompt (Figure 1) a set of potential additions and returns a subset to undergo human review (second stage). The steps are as follows:

1. Identify the classes to be supplemented with new comments, $\mathcal{M}$.
2. For each $m \in \mathcal{M}$, do
2.1. For each comment c_m in the class m, do
2.1.1. Identify *five* nearest neighbours (using the embedding space[14]) of the comment, c_{mn}, from the unlabelled portion of the dataset.
2.1.2. Pass the comments c_{mn} in the prompt (Figure 1 provides the structure) as candidates for the LLM to choose from as relevant to the class m.
2.1.3 The LLM returns a subset of the candidate comments, $c_{llm} \subseteq c_{mn}$. The collection of LLM-returned comments for m is indicated by C_{llm}

The Second Stage: Human Approval.

Here, human annotators review (and not annotate) the LLM-returned comments.
2.2. The LLM-returned comments C_{llm} are reviewed by human annotators for correct comment-class association.
2.3. The unanimously approved comments, $C_{approved} \in C_{llm}$, are incorporated into the already annotated dataset, $C_{approved} \rightarrow \mathcal{D}$. Repeat for each m.

[14] We generate the embeddings using Google's embedding model `embedding-004`.

As shown in Figure 1, we utilize a *five* segment prompt with nearest neighbours of each anchor comment as candidates for potential addition, which were otherwise to be reviewed by humans. To maintain equal representation among *four* stance-indicating classes not considering *Undetermined*, we choose to upscale the classes that contribute most to the imbalance due to a low number of comments – for *Triple Talaq* and *Section 377*: Direct Critic, Indirect Critic, Direct Favor; for *Sabarimala*: Direct Critic, Direct Favor, Indirect Favor. To deploy the approach described above, we utilize the Gemini-Pro 1.5 LLM for its ability of contextual grounding [46]. In the attempt to minimize manual efforts, utilization of LLM reduces the total number of neighbours to be reviewed by the human annotators in the second stage down by 42.46% in *Triple Talaq*, 28.4% in *Section 377* and 46.41% in *Sabarimala*. LLM here reduces the efforts since *only* LLM-returned comments are only *reviewed* in the second stage.

Table 4 shows the previous and revised distributions, produced by the two-stage approach. After the deployment of the approach, there is a noticeable difference in the distributions. While *Triple Talaq* has, although low in number but mostly positive comments reflecting majority support for the change, the *Sabarimala* verdict finds more criticism. Therefore, the number of critic comments in *Triple Talaq* is substantially low and for *Sabarimala* the exact opposite is true. *Section 377* however shows the approach can be effective if each class has an adequate number of comments in the unlabelled portion of the dataset.

While a manual inspection of the class distributions in Table 4 suggests that our approach has reduced class imbalance, in what follows, we provide a quantitative analysis of imbalance. Following a similar approach as Ansari *et al.* [1], we formalize a quantifiable measure of class imbalance. Consider a dataset $\mathcal{D}$ with k classes denoted by $\mathcal{C}_1$, $\mathcal{C}_2$, ..., $\mathcal{C}_k$. Consider $\mathcal{C}_i$ represents l_i instances in $\mathcal{D}$.

We first construct the probability vector $\mathbf{l}$:

$$\left[\frac{l_1}{\sum_{i=1}^{k} l_i}, \frac{l_2}{\sum_{i=1}^{k} l_i}, \ldots, \frac{l_k}{\sum_{i=1}^{k} l_i}\right] \tag{1}$$

In simple words, the j-th element of this probability vector $\mathbf{l}$ denotes the fraction of class $\mathcal{C}_j$ instances in $\mathcal{D}$.

For $\mathcal{D}$, we define the imbalance as the KL divergence between this probability vector and a uniform discrete probability vector with k elements.

$$\mathit{imbalance}(\mathcal{D}) = KL(\mathbf{l}, [\tfrac{1}{k}, \tfrac{1}{k}, \ldots, \tfrac{1}{k}]) \tag{2}$$

$\mathit{imbalance}(\mathcal{D})$ measures how far the class distribution of $\mathcal{D}$ is from a uniform discrete distribution, i.e., a perfectly balanced scenario where each class has equal representation in the dataset. If $\mathcal{D}$ is perfectly balanced, $\mathit{imbalance}(\mathcal{D})$ will be 0. A higher value of this measure indicates greater imbalance. As shown in Table 6, among the three datasets, *Triple Talaq* exhibits the highest imbalance value, suggesting the most uneven distribution. We further observe considerable reduction in imbalance after employing our human-LLM collaborative approach.

	Triple Talaq		Section 377		Sabarimala	
	Accuracy	mac-F1-Score	Accuracy	mac-F1-Score	Accuracy	mac-F1-Score
Mistral (Mistral-7B-Instruct-v0.3)	77.23±1.50	65.76±2.19	84.68±0.83	85.43±0.94	79.79±0.64	78.92±0.81
Llama3 (Llama-3.1-8B-Instruct)	80.67±0.90	68.67±3.99	83.13±0.81	84.34±0.53	77.26±1.58	76.34±1.33
DeepSeek (DeepSeek-R1-Distill-Llama-8B)	75.36±2.73	59.62±2.69	80.21±1.37	80.97±1.42	73.24±1.92	69.55±2.06
SVM (Google Embeddings)	79.17±1.55	65.19±5.69	82.21±2.76	82.00±2.55	75.28±2.19	68.27±3.29
NB (Google Embeddings)	72.88±3.21	59.35±7.44	78.49±2.82	79.5±3.41	71.33±1.75	67.76±2.06
BERT-Base-Uncased	70.32±3.66	57.54±4.73	75.60±1.47	76.83±1.61	70.44±2.84	65.83±2.91
DistilBERT	71.26±2.79	53.23±2.10	74.50±2.68	75.86±1.94	70.17±2.88	66.56±2.71
RoBERTa	72.69±3.49	55.24±5.00	75.40±1.36	76.96±1.20	70.69±2.11	65.96±1.89
`Gemini`$_{zero}$	67.28±1.02	55.92±1.36	77.97±3.15	73.00±3.86	71.39±3.16	71.52±0.96
`Llama3.1-8B`$_{zero}$	48.81±0.61	44.05±0.56	70.26±3.43	71.57±2.62	63.89±4.05	65.96±3.40

Table 5: Classifier perfomrance on our datasets. Zero-shot models (`Llama3.1` and `Gemini-1.5`) are indicated with a subscript *zero*. For each dataset, the best classifier performance is highlighted.

Dataset	Before	After	Change in imabalance
Sabarimala	0.49	0.32	-0.17
Section 377	0.24	0.13	-0.11
Triple Talaq	0.79	0.69	-0.10

Table 6: Dataset imbalance as computed using Equation 2. A higher value indicates greater imbalance. The *before* column lists the imbalance in individual datasets before we employed our human-LLM collaboration pipeline to address imbalance. The *after* column lists the imbalance in individual classes after we addressed class imbalance as described in Section 2.3. We observe that for all datasets, our approach has considerably reduced class imbalance.

3 Results and Analyses

3.1 Stance Classifier Performance

We consider three well-known open large language models fine-tuned on our datasets: `Llama3` [13]; `Mistral` [23]; and `Deepseek` [5]. In addition, following prior literature in stance classifiers (e.g., [22, 28, 37, 45]), we consider Support Vector Machine (SVM) [8]; Naive Bayes (NB) [6], and three `BERT`-based models (`BERT` [11]; `DistilBERT` [44]; and `RoBERTa` [30]). We use 70:30 train/test splits and for each model. Refer to Supplemental Information (SI) for training details.

Following standard machine learning practices for evaluating classifiers on imbalanced datasets [6], we use the macro-F1 score as our primary performance evaluation metric. In addition, in Table 5, we report model accuracy on individual datasets. As shown in Table 5, we observe that (1) fine-tuned models perform considerably better than the zero-shot models; and (2) compared to all other baselines `Mistral` and `LLaMA 3.1` perform the best with `Mistral` achieving best performance on two datasets (*section 377* and *Sabarimala*) and `LLaMA 3.1` achieving best performance on the remaining one (*Triple Talaq*). Across all three datasets, the best performance is achieved on *Section 377*. We note that, this dataset has the lowest imbalance as shown in Table 6.

Error Analysis: We observe that the models mostly struggled with distinguishing between indirect stances (favor or critic) and undetermined. Upon manual

inspection, we observe a recurring theme. There were instances where the models could not understand nuanced cultural contexts. For instance, the comment *As usual Chee News* is a wordplay on Zee News (the news outlet) and the Hindi word `Chee` which is used to express disgust. The model classified this example as undetermined while the annotators labeled it as indirect critic. Similarly, the comment *Congrats to those who wanted to see INDIA like AMERICA ..* was annotated as Indirect Critic while the models predicted it as Undetermined. For a sizable conservative population in India, America (and the West in general) is looked down upon as a country with a highly visible queer population, which conservative Indians believe has destroyed the cultural fabric there. SI contains a detailed exposition of error analysis with confusion matrices.

Classes	**Triple Talaq**	**Section 377**	**Sabarimala**
Direct Critic	0.19%	2.64%	4.23%
Indirect Critic	3.94%	21.02%	33.5%
Direct Favor	3.11%	4.59%	1.41%
Indirect Favor	17.79%	14.60%	4.59%
Undetermined	74.95%	57.11%	56.21%

Table 7: Distributions of the remaining i.e. unlabelled comments on deployment of the best models.

Triple Talaq
▶ Brought tears, even I am also victim of triple talaq, I am an well educated gal alhmdlh but cudn't raise my voice against injustice fr which I regrate myself, I just want to be like u strong enough to fight against sch evils
Section 377
▶ My goodness, we finally got the freedom to live, 24 years of struggle and i cant stop crying Congrats India
Sabarimala
▶ Sati pratha was also a tradition and what if it wasn't stopped then would people still doubt that it is our thousand year old culture and no one has right to interfere? I am completely aware about the tradition and culture with sabarimala but this is 21st century and suppression of any human being Wether men or women needs to be stopped...

Table 8: Comments indicating a favorable stance found in the wild. Refer to SI for more such comments.

3.2 Discourse Around the Verdicts

The best-performing model per verdict is run on the unlabeled portion of the corresponding dataset Table 7 summarizes the distributions of comments in the wild according to the model. We first note that not all verdicts received a similar share of support or criticism – civic receptivity considerably varied across different judgments. Among all three verdicts, *Sabarimala* received minimal support while *Triple Talaq* received the maximum support (see, Figure 2). The overwhelming support for *Triple Talaq* aligns with the fact that more than 1 million Muslims in India signed a petition to abolish Triple Talaq. Table 8 shows an example comment in which the author self-identifies as a victim of this law.

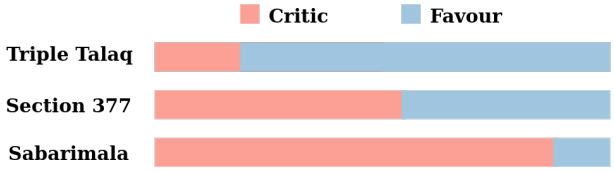

Fig. 2: Distributions of the unlabelled comments indicate the tilt in polarity is overwhelmingly negative for *Sabarimala* and positive for *Triple Talaq* while for *Section 377* it is comparatively balanced.

As highlighted already, our dataset is annotated by a self-identified queer person from India. Before homosexuality was decriminalized in India, being openly gay in India was an uphill task. Beyond many comments reiterating *love is love*, we observe a rare humanization of a particularly invisible community in India where several queer people voiced support in the first person in the comments section. Finally, in the discourse around *Sabarimala*, we notice several comments drawing parallel with this landmark judgment with the abolishment of Sati by Raja Ram Mohun Roy in 1829. Table 8 lists example comments.

4 Conclusion

In the study, we consider three landmark judgments of the Indian Supreme Court that encompass two major aspects of Indian society – gender and religion and, additionally, sexual identity, in alignment with the Sustainable Development Goal of achieving equality. We observe that verdicts around disadvantaged groups receive varied responses from the public, with the ban on instantaneous *Triple Talaq* receiving a major upvote and the entry of menstruating women at Sabarimala being highly criticized. We look forward to considering comments in other languages especially regional Indian languages and extending our study to other jurisdictions.

Limitations of the study

For the study, we consider a single platform (YouTube) and a monolingual setup – only English, which constitutes around 50% of the collection (Table 2). Due to annotator unavailability, our work does not consider regional languages. However, the verdicts triggered widespread pan-India discussions which were mostly held in English. India's demography provides a high degree of varied opinions, which we believe provides substantial sample for an empirical study. The identified limitations highlight an opportunity for future research to include multilingual data across multiple platforms for a more holistic analysis.

Ethical Statement

We have used only public data on the Web, collected using the publicly available YouTube API. We did not collect any user identities or user-details during the data collection process.

The LLMs used may contain inherent biases. We have compensated the annotators commensurate with their efforts, and the annotators were informed of the purpose of their annotations.

Given the sensitive nature of the data, we have anonymized the texts of the comments in our datasets.

Acknowledgement

Koustav Rudra is a recipient of the DST-INSPIRE Faculty Fellowship (DST/INSPIRE/04/2021/003055 in the year 2021 under Engineering Sciences) We acknowledge the annotators who helped us in developing the dataset.

References

1. Ansari, M.A., Sidhpura, J., Mandal, V.K., Khudabukhsh, A.R.: Quantifying the transience of social web datasets. In: ASONAM. pp. 286–293 (2023)
2. Ash, E., Asher, S., Bhowmick, A., Bhupatiraju, S., Chen, D.L., Devi, T., Goessmann, C., Novosad, P., Siddiqi, B.: In-group bias in the Indian judiciary: Evidence from 5 million criminal cases. Tech. rep., Center for Global Development (2023)
3. Ayesha Jamal: Sabarimala Verdict: A Watershed Moment in the History of Affirmative Action (20 October 2020)
4. BBC: Triple talaq: India criminalises Muslim 'instant divorce' (30 July 2019), https://www.bbc.com/news/world-asia-india-49160818
5. Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al.: Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954 (2024)
6. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg (2006)
7. Bondi, E., Xu, L., Acosta-Navas, D., Killian, J.A.: Envisioning communities: a participatory approach towards AI for social good. In: AIES 2021. pp. 425–436 (2021)
8. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. pp. 144–152 (1992)
9. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: ICWSM 2017. pp. 512–515 (2017)
10. Delgado, F., Barocas, S., Levy, K.: An uncommon task: Participatory design in legal AI. CSCW 6(CSCW1), 1–23 (2022)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL: HLT. pp. 4171–4186. ACL (2019), https://aclanthology.org/N19-1423

12. Dodwell, H.: The Cambridge history of the British empire, vol. 5. CUP Archive (1932)
13. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
14. Dutta, S., Srivastava, P., Solunke, V., Nath, S., KhudaBukhsh, A.R.: Disentangling societal inequality from model biases: Gender inequality in divorce court proceedings. In: IJCAI 2023. pp. 5959–5967 (2023)
15. Gao, L., Huang, R.: Detecting online hate speech using context aware models. arXiv preprint arXiv:1710.07395 (2017)
16. Gautam, A., Mathur, P., Gosangi, R., Mahata, D., Sawhney, R., Shah, R.R.: # metooma: Multi-aspect annotations of tweets related to the metoo movement. In: ICWSM. pp. 209–216 (2020)
17. Glandt, K., Khanal, S., Li, Y., Caragea, D., Caragea, C.: Stance detection in covid-19 tweets. In: ACL-IJNLP 2021. vol. 1 (2021)
18. Grasso, F., Locci, S., Siragusa, G., Caro, L.D.: Ecoverse: An annotated twitter dataset for eco-relevance classification, environmental impact analysis, and stance detection (2024), https://api.semanticscholar.org/CorpusID:269004840
19. Gyawali, N., Sirbu, I., Sosea, T., Khanal, S., Caragea, D., Rebedea, T., Caragea, C.: Gunstance: Stance detection for gun control and gun regulation. In: ACL 2024. pp. 12027–12044 (2024)
20. Harrington, C., Erete, S., Piper, A.M.: Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. CSCW 3(CSCW), 1–25 (2019)
21. He, X., Lin, Z., Gong, Y., Jin, A.L., Zhang, H., Lin, C., Jiao, J., Yiu, S.M., Duan, N., Chen, W.: AnnoLLM: Making large language models to be better crowdsourced annotators. In: NAACL: HLT (Volume 6: Industry Track). pp. 165–190. ACL (Jun 2024), https://aclanthology.org/2024.naacl-industry.15/
22. He, Z., Mokhberian, N., Lerman, K.: Infusing knowledge from wikipedia to enhance stance detection. arXiv preprint arXiv:2204.03839 (2022)
23. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
24. Joshi, A., Paul, S., Sharma, A., Goyal, P., Ghosh, S., Modi, A.: IL-TUR: Benchmark for Indian Legal Text Understanding and Reasoning. In: ACL 2024. pp. 11460–11499 (2024)
25. Khandkar, S.H.: Open coding. University of Calgary 23(2009) (2009)
26. Kholodna, N., Julka, S., Khodadadi, M., Gumus, M.N., Granitzer, M.: Llms innbsp;thenbsp;loop: Leveraging large language model annotations fornbsp;active learning innbsp;low-resource languages. In: ECML PKDD 2024. p. 397–412. Springer-Verlag, Berlin, Heidelberg (2024), https://doi.org/10.1007/978-3-031-70381-2_25
27. Khorramrouz, A., Dutta, S., KhudaBukhsh, A.R.: For Women, Life, Freedom: A Participatory AI-Based Social Web Analysis of a Watershed Moment in Iran's Gender Struggles. In: IJCAI 2023. pp. 6013–6021 (2023)
28. Lan, X., Gao, C., Jin, D., Li, Y.: Stance detection with collaborative role-infused llm-based agents. In: ICWSM. vol. 18, pp. 891–903 (2024)
29. Li, Y., Zhang, Y.: Pro-woman, anti-man? identifying gender bias in stance detection. In: Findings of ACL 2024. pp. 3229–3236 (2024)

30. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019), https://arxiv.org/abs/1907.11692
31. Malik, V., Sanjay, R., Nigam, S.K., Ghosh, K., Guha, S.K., Bhattacharya, A., Modi, A.: ILDC for CJPE: indian legal documents corpus for court judgment prediction and explanation. In: ACL/IJCNLP 2021. pp. 4046–4062 (2021)
32. Mani, L.: Contentious traditions: The debate on sati in colonial India. Univ of California Press (1998)
33. Mitra, D.: History's apology: Sexuality and the 377 supreme court decision in india. Epicenter, Harvard University (27 September 2018)
34. Nigam, S.K., Patnaik, B.D., Mishra, S., Shallum, N., Ghosh, K., Bhattacharya, A.: NYAYAANUMANA and INLEGALLLAMA: the largest indian legal judgment prediction dataset and specialized language model for enhanced decision analysis. In: COLING 2025. pp. 11135–11160. ACL (2025), https://aclanthology.org/2025.coling-main.738/
35. Nigam, S.K., Sharma, A., Khanna, D., Shallum, N., Ghosh, K., Bhattacharya, A.: Legal Judgment Reimagined: PredEx and the Rise of Intelligent AI Interpretation in Indian Courts. In: Findings of ACL. pp. 4296–4315 (2024)
36. Palakodety, S., KhudaBukhsh, A.R., Carbonell, J.G.: Hope speech detection: A computational analysis of the voice of peace. In: ECAI 2020. pp. 1881–1889 (2020)
37. Palakodety, S., KhudaBukhsh, A.R., Carbonell, J.G.: Voice for the voiceless: Active sampling to detect comments supporting the rohingyas. AAAI 2020 34(01), 454–462 (Apr 2020)
38. Pangakis, N., Wolken, S., Fasching, N.: Automated annotation with generative ai requires validation. ArXiv abs/2306.00176 (2023), https://api.semanticscholar.org/CorpusID:259000016
39. Paul, S., Bhatt, R., Goyal, P., Ghosh, S.: Legal statute identification: A case study using state-of-the-art datasets and methods. In: SIGIR 2024. pp. 2231–2240. ACM (2024)
40. PTI: British era survey report says Sabarimala ban existed 200 years ago. The Week (22 November 2018)
41. QueerInAI, O.O., Ovalle, A., Subramonian, A., Singh, A., Voelcker, C., Sutherland, D.J., Locatelli, D., Breznik, E., Klubicka, F., Yuan, H., Jethwani, H., et al.: Queer in AI: A case study in community-led participatory AI. In: FAccT 2023. pp. 1882–1895. ACM (2023), https://doi.org/10.1145/3593013.3594134
42. Rao, R.: Out of time: The queer politics of postcoloniality. Oxford University Press pp. 7–9 (2020)
43. Reiss, M.V.: Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. ArXiv abs/2304.11085 (2023), https://api.semanticscholar.org/CorpusID:258291402
44. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (2020)
45. Siddiqua, U.A., Chy, A.N., Aono, M.: Stance detection on microblog focusing on syntactic tree representation. In: DMBD 2018. pp. 478–490. Springer (2018)
46. Team, G Gemini: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. URL https://goo. gle/GeminiV1-5 (2024)
47. The Economist: Recent court rulings in India suggest justice is improving (31 August 2017)
48. Wiegand, M., Ruppenhofer, J., Kleinbauer, T.: Detection of abusive language: the problem of biased datasets. In: NAACL-HLT 2019. pp. 602–608 (2019)