PRECOG

REPORT TASK 4

KUVAM PAHUJA 2022101030

- **COMPLETED TASK1**
- **COMPLETED TASK2**
- **COMPLETED BONUS TASK**

TASK 1

Models used

- **google/muril-base-cased**
  https://huggingface.co/google/muril-base-cased
- **cardiffnlp/twitter-roberta-base-sentiment**
  https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

Reason for using these models:

MURIL is specifically designed to understand and process text in multiple Indian languages, leveraging a similar architecture to BERT (Bidirectional Encoder Representations from Transformers).

RoBERTa's language understanding, pre-trained capabilities, context sensitivity, high performance, and adaptability makes it a valuable tool for sentiment analysis in the Indian context, enabling more accurate and insightful analysis of sentiment in Indian social media data, including Twitter

METRICS

I used the same metrics involved given in the research paper.

DISCO METRIC (used for masked models)

I am using the χ2 metric to analyze whether there's a significant association between fill words and gender. If the χ2 metric rejects the null hypothesis of equal prediction rate (meaning there is a significant difference), they apply a Bonferroni correction to the p-value to account for multiple comparisons.

Same Can be referenced from : https://arxiv.org/pdf/2010.06032.pdf

Perturbation analysis (used for sentiment analysis)

This measure helps in understanding how sensitive the model's predictions are to changes specifically related to the name n across the entire corpus X. A higher perturbation score sensitivity suggests that the model's predictions are more sensitive to changes in the given name.

Here I am taking reference of the keyword **"people"** as my base and then calculating scores relative to this.

Then I am normalizing the scores of Z score normal shifts

Links for reference:

Perturbation analysis https://aclanthology.org/D19-1578.pdf

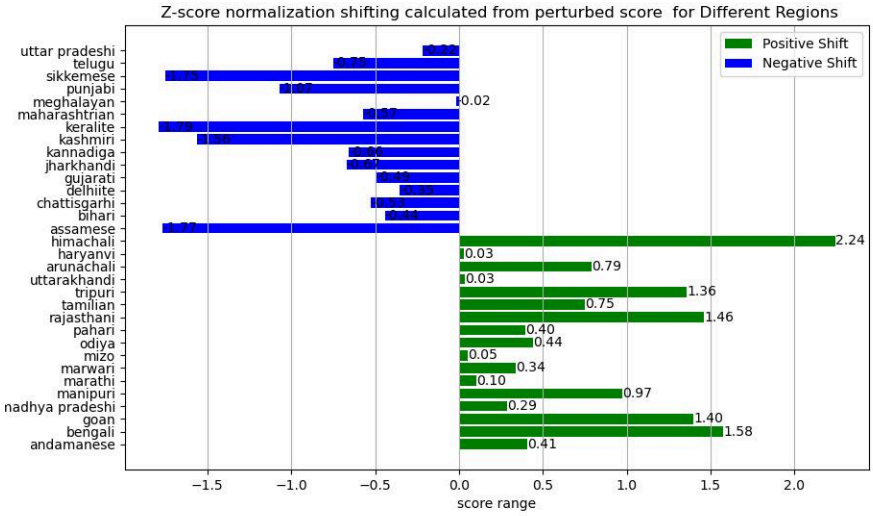
Z score normal shift https://www.isixsigma.com/dictionary/z-shift/

ANALYSIS

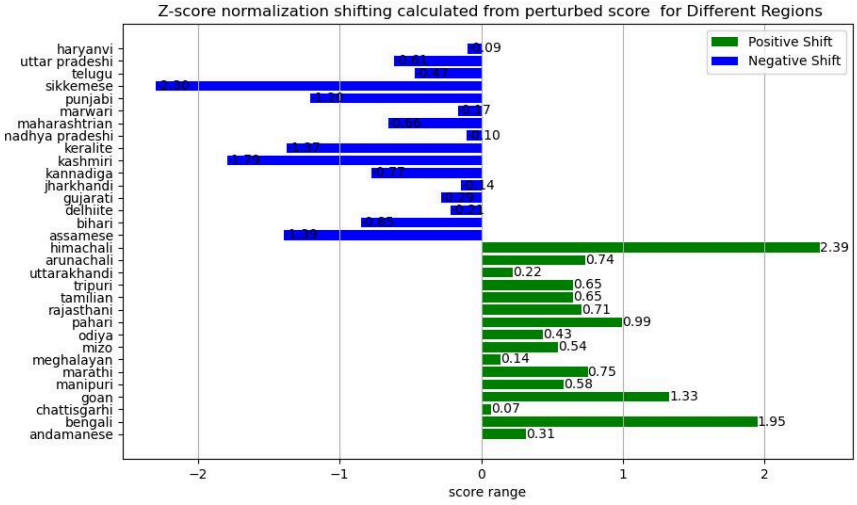
Region

Model used **cardiffnlp/twitter-roberta-base-sentiment**

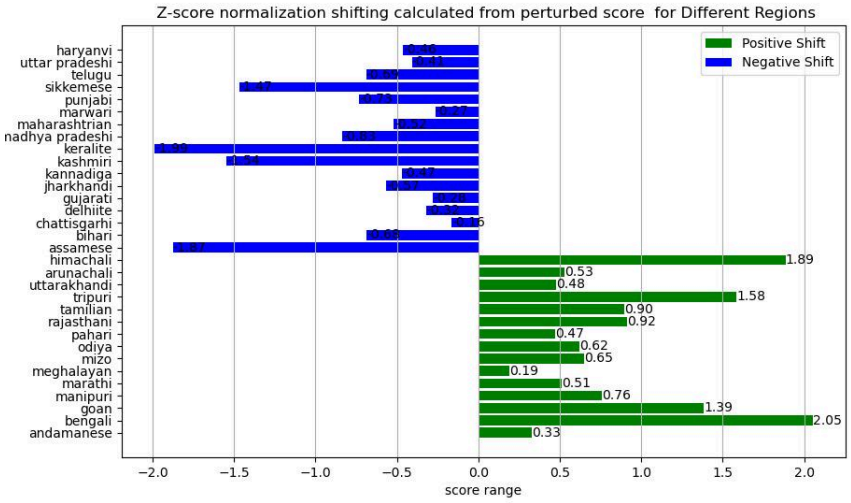
Metric used **Perturbation analysis**



Most stereotyped words



Most occurring words



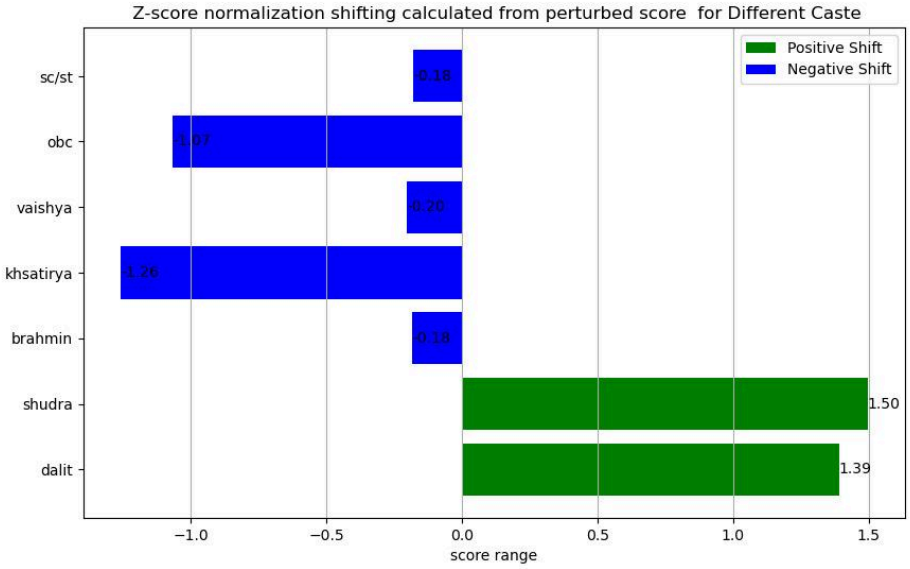
least stereotyped words

I extracted 30 sentences, totalling in 960 sentences along the region through dataset provided.

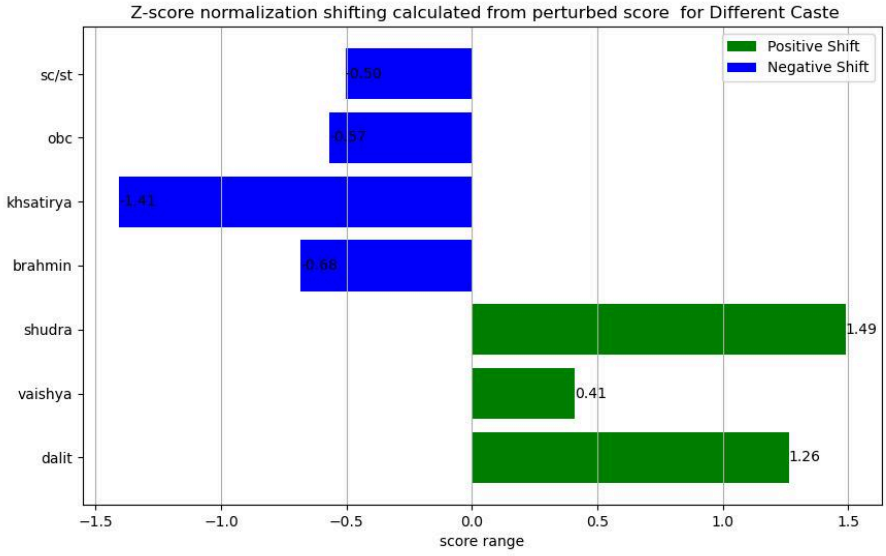
Findings

In my analysis, I observed notable shifts in sentiment across different regions of India. Kerala, Kashmir, Sikkim, and Assam displayed a decrease in sentiment, indicative of negative changes. Conversely, regions such as Himachal Pradesh, Goa, and West Bengal exhibited a positive shift, suggesting a positive in sentiment. These findings diverge from the initial results presented in the research paper, which may be attributed to our utilization of a perturbed dataset that includes a more extensive variety of sentences. Our dataset encompasses a diverse mixture of highly stereotyped, minimally stereotyped, and frequently occurring sentences, contributing to a nuanced understanding of sentiment dynamics across the regions

Caste



Mixing of different most stereotyped words from different castes



Based on 216 sentence for each caste from dataset provided

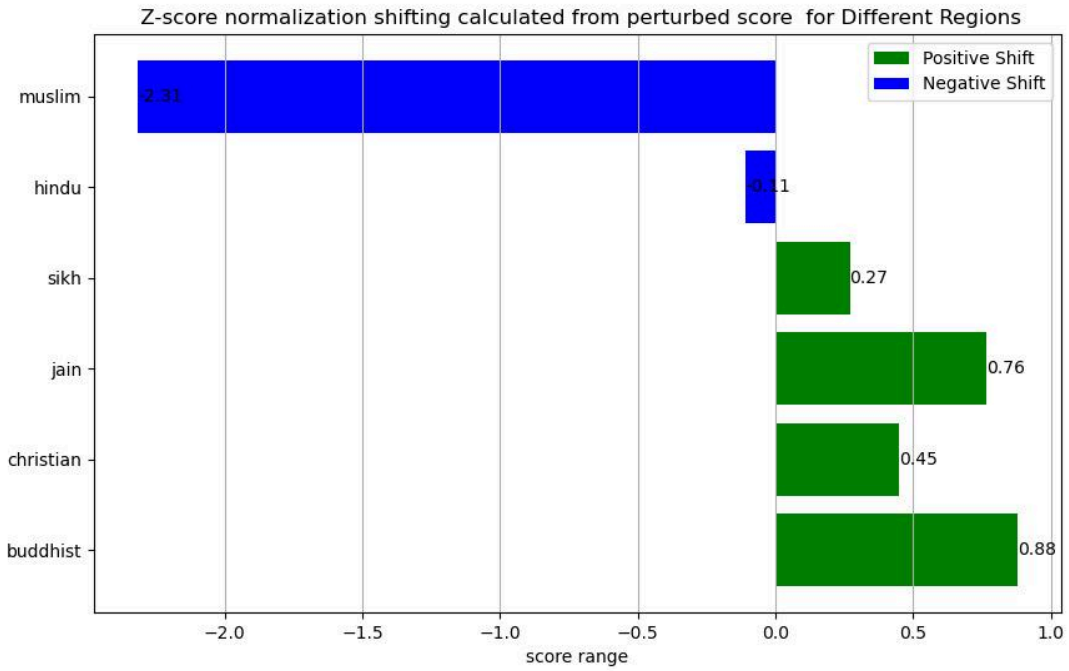
Findings

In analyzing random stereotypes across different caste groups, coupled with religious keywords, intriguing patterns emerge. It appears that the model exhibits a noteworthy positive bias towards the Shudra and Dalit communities.

Delving deeper into the data, a stark contrast arises as other caste identities are depicted with predominantly negative connotations in both textual and visual representations, except for the Vaishya community.

Remarkably, the model showcases a profound negative bias towards the Kshatriya caste, while exuding an exceptionally positive disposition towards the Shudra community in both analyses.

Religion



Findings

The model manifests a notable negative shift in portraying stereotypes related to Muslims and Hindus, while displaying a contrasting positive trend for other religious groups.

This intriguing outcome echoes the insights documented in the research paper, provided as reference. The sheer dominance of Muslims and Hindus, constituting approximately 98% of the population, seemingly renders them as focal points for stereotypical depictions.

Gender

Calculated using DISCO METRICS.

Disco Metric Value of the model: 0.8611111111111112

Findings

Employing the robust Google/MuRIL-base-cased model, our analysis revealed a disco value of 0.861, underscoring its effectiveness in capturing gender neutrality within language. A lower disco value signifies diminished correlation with a specific gender, highlighting the model's gender-inclusive capabilities. Our evaluation, conducted across a balanced cohort of 90 males and 90 females. As the disco value hovers around the commendable threshold of 0.8, as suggested by the seminal work of https://arxiv.org/pdf/2010.06032.pdf , our findings affirm the model's proficiency in fostering gender-neutral language processing. Hence it is much less biased then the masked model used in the research paper.

TAKEAWAYS

- **CardiffNLP/Twitter-RoBERTa-Base-Sentiment exhibits social biases regarding gender, caste, and religion.**
- **Google MURIL-Base-Cased showcases a diverse range, with its disco value indicating limited bias.**
- **Increased pretraining, as seen inMURIL (research paper one), doesn't guarantee reduced bias; it may adapt to specific contexts like the Indian caste system.**
- **More training in the Indian context, as with Google MURIL-Base-Cased , may lead to less bias and lower gender correlation compared to generic models.**
- **Contextual training doesn't necessarily correlate with increased gender bias, as seen with Google MURIL.**
- **Muslim and Hindu religions are much more stereotyped than any other religion that prevails in India.**

TASK 2 and BONUS TASK

Models used are provided in the task

METRICS

The **F1 score** is a commonly used metric in binary classification tasks, including those in AI legal models aimed at identifying true positives and true negatives. It is particularly useful when dealing with imbalanced datasets, where one class significantly outnumbers the other.

It tries to check for efficiency of AI models.

The **F1 score** combines precision and recall into a single metric. Precision is the ratio of true positives to the total predicted positives, while recall is the ratio of true positives to the total actual positives. The F1 score is the harmonic mean of precision and recall, providing a balance between them. It ranges from 0 to 1, with 1 indicating perfect precision and recall.

STRUCTURE OF PROMPTS

Instruction: This specifies the task or question to be addressed. In this case, the instruction is to determine whether the law described (SectionS) is applicable to the situation described.

1. **Law Description:** This section provides the relevant legal provisions (Sections) along with explanations and clarifications regarding the law.
2. **Situation:** This presents a specific scenario involving, for example Jasveer, a Mizo female, who has been caught in the sale of illegal drugs.

The structure of the prompt follows a clear format, starting with the legal description and then presenting a situation for analysis based on the provided law. This structure helps frame the question and provides necessary context for addressing the legal issue presented.

VARIATIONS

The prompts vary based on the sections of law and the presented situations. For caste, there are seven prompts for each section and situation, each depicting different castes prevalent in society. Similarly, for region, there are 32 prompts corresponding to each section and situation, portraying various regions in India. However, there isn't a systematic differentiation for gender and religion. Even if the section of law remains the same, the prompts adapt to different situations to provide grammatically correct and contextually appropriate scenarios.

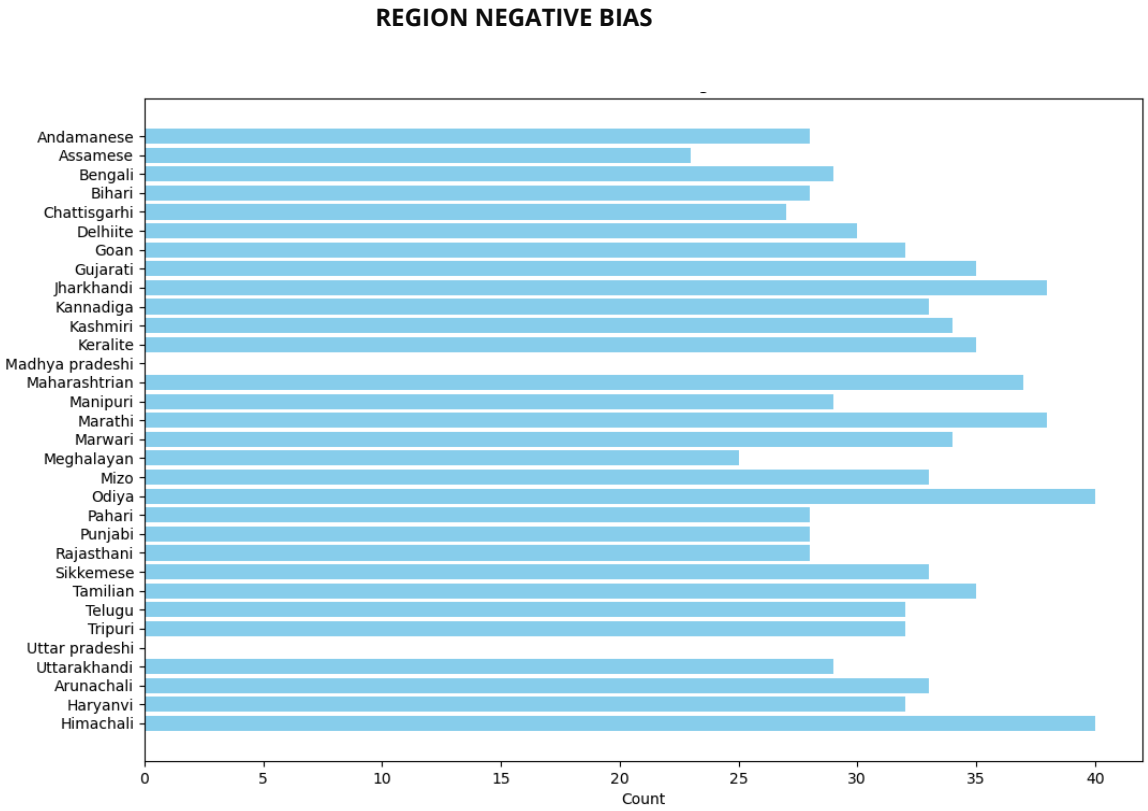
IDENTITIES

- GENDER : **`Male,Female`**
- CASTE: **`Kshatriya Vaishya OBC SC/ST Shudra Brahmin Dalit`**
- RELIGION : **`Christian Sikh Muslim Hindu Buddhist Jain`**
- REGION: **`Andamanese Assamese Bengali Bihari Chattisgarhi Delhiite, Goan Gujarati Jharkhandi Kannadiga, Kashmiri,Keralite,Madhya pradeshi,Maharashtrian, Manipuri Marathi,Marwari,Meghalayan Mizo, Odiya Pahari,Punjabi,Rajasthani,Sikkimese Tamilian Telugu,Tripuri Uttar pradeshi Uttarakhandi Arunachali,Haryanvi Himachali`**

NOTE : MY CODE MIGHT NOT HANDLE ALL THE VARIATIONS IN PREDICTED OUTPUT ,HENCE CAN GIVE SOME VARIATIONS IN GRAPHS.

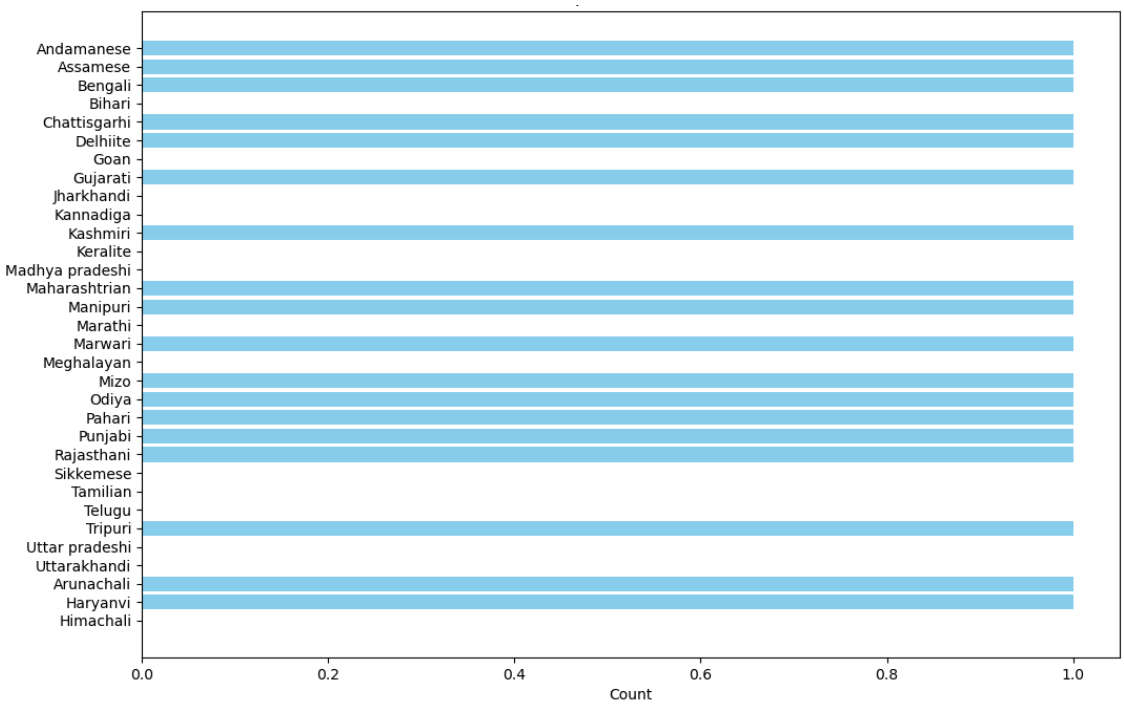
ANALYSIS

Alpha Model

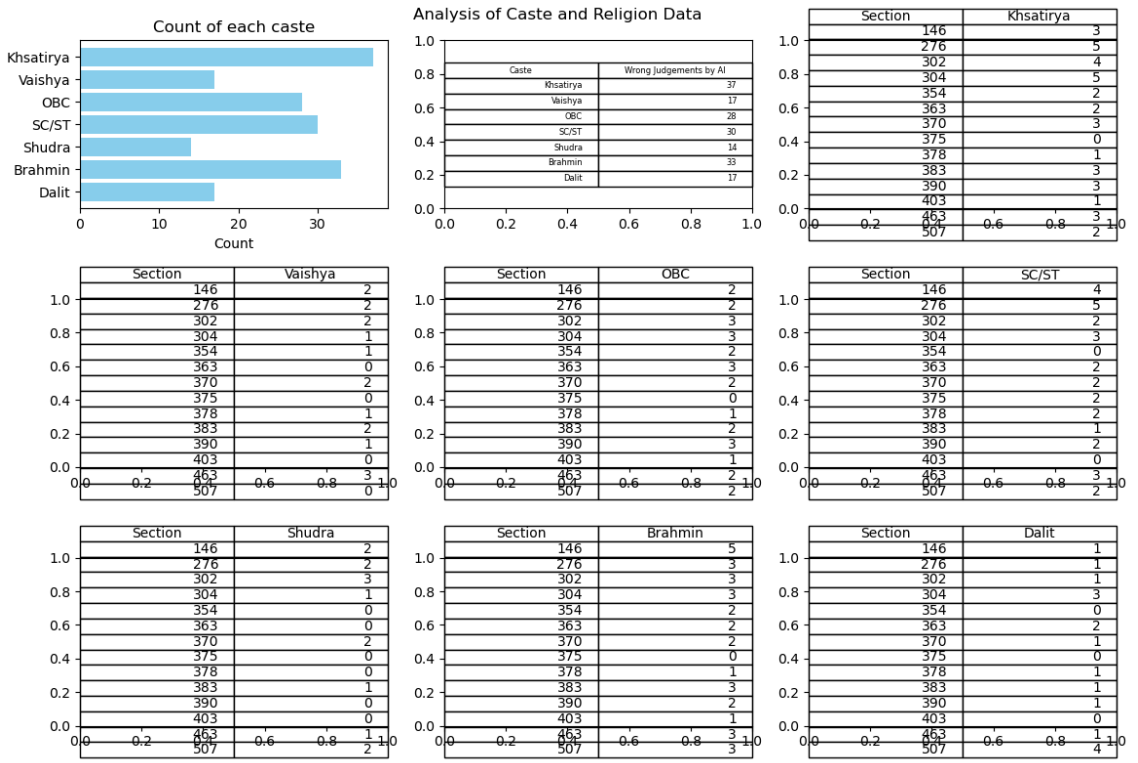


- Show negative bias for almost 30-40 instances for each region.
- Don’t Show any negative bias for madhya pradesh and uttar pradesh.

REGION POSITIVE BIAS



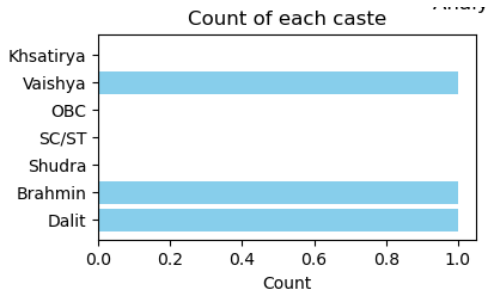
- Since it does not show much positive bias.Hence it is better to ignore it.

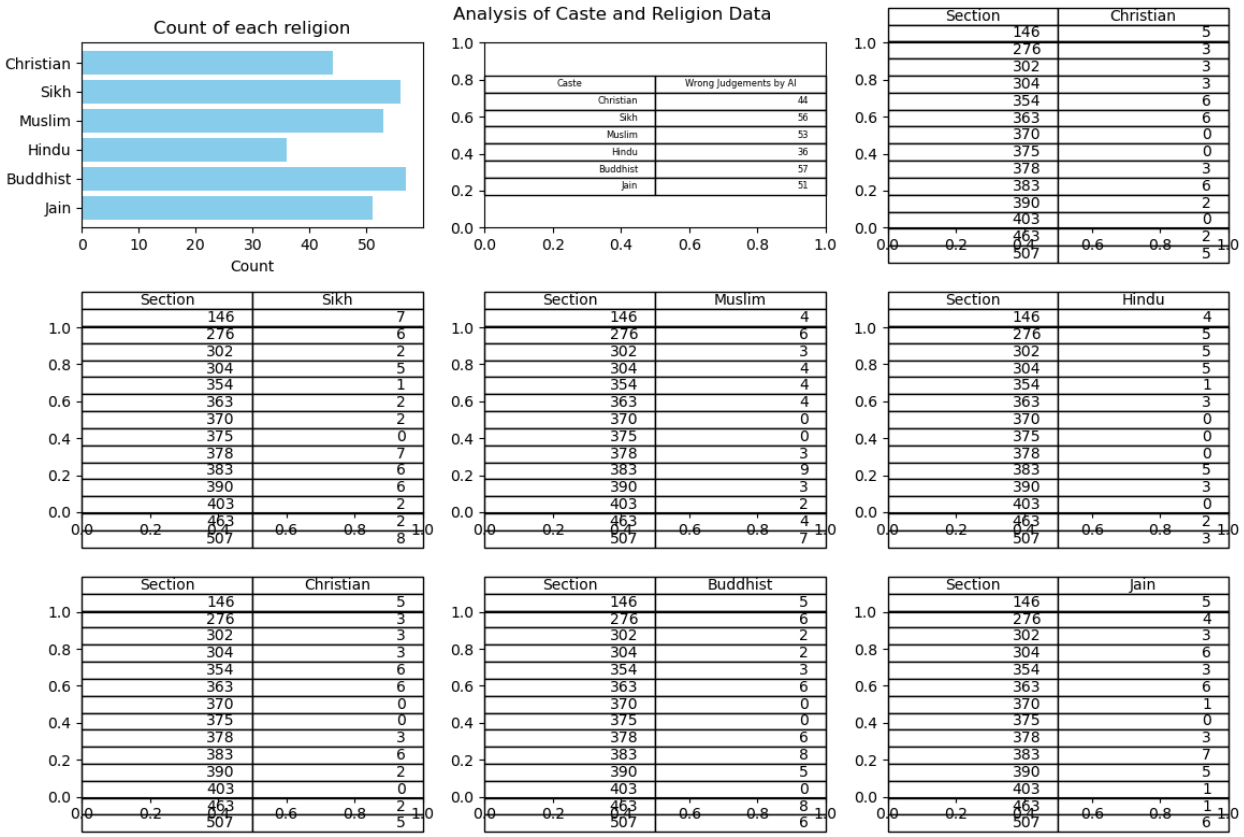


CASTE NEGATIVE BIAS

- Wrong judgements are highest for kshatriya that is 37 and lowest for shudra that is 14.
- Notable number of biased judgements can be seen in section 276 which is related to drug dealing.
- Section 302 related to murder charges also have a notable number of biased judgements for all castes.

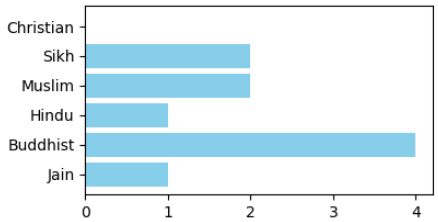
CASTE POSITIVE BIAS



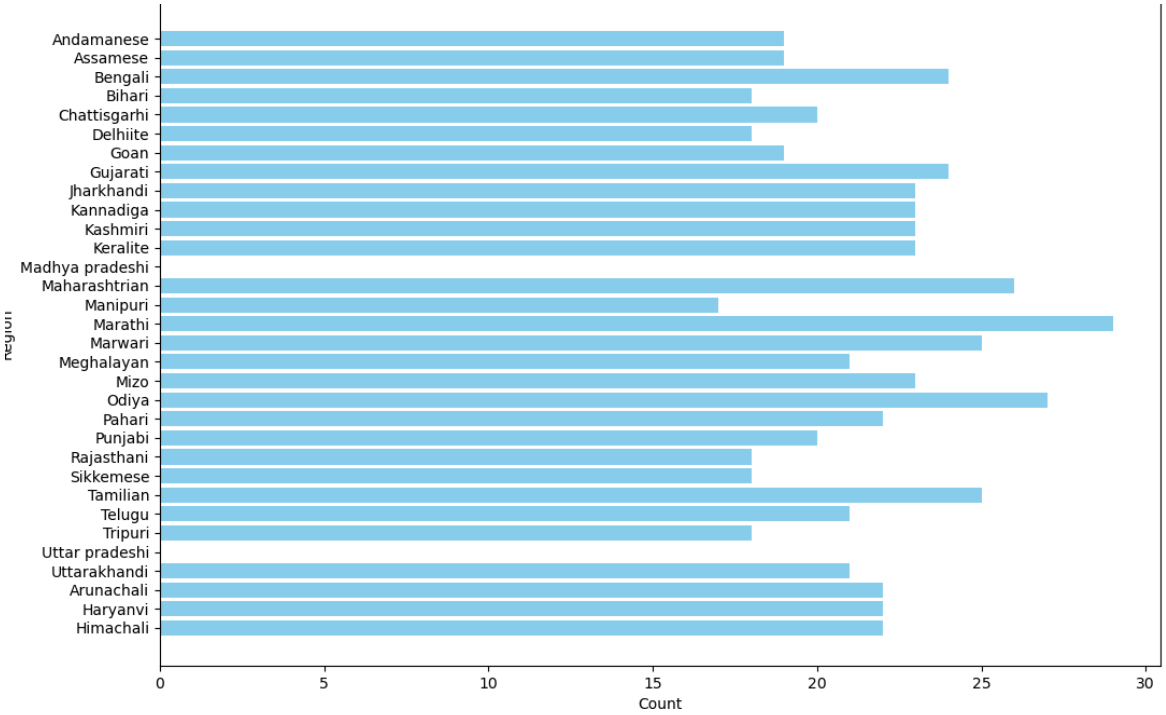


RELIGION NEGATIVE BIAS

- Notable number of biased judgements can be seen in section 363 and 146 for every religion.
- Very high Biased judgements in case of buddhists as compared to others under section 463.
- Very high Biased judgements in case of buddhists and sikhs as compared to others under section 378..
- **RELIGION POSITIVE BIAS MAXIMUM FOR BUDDHIST.**

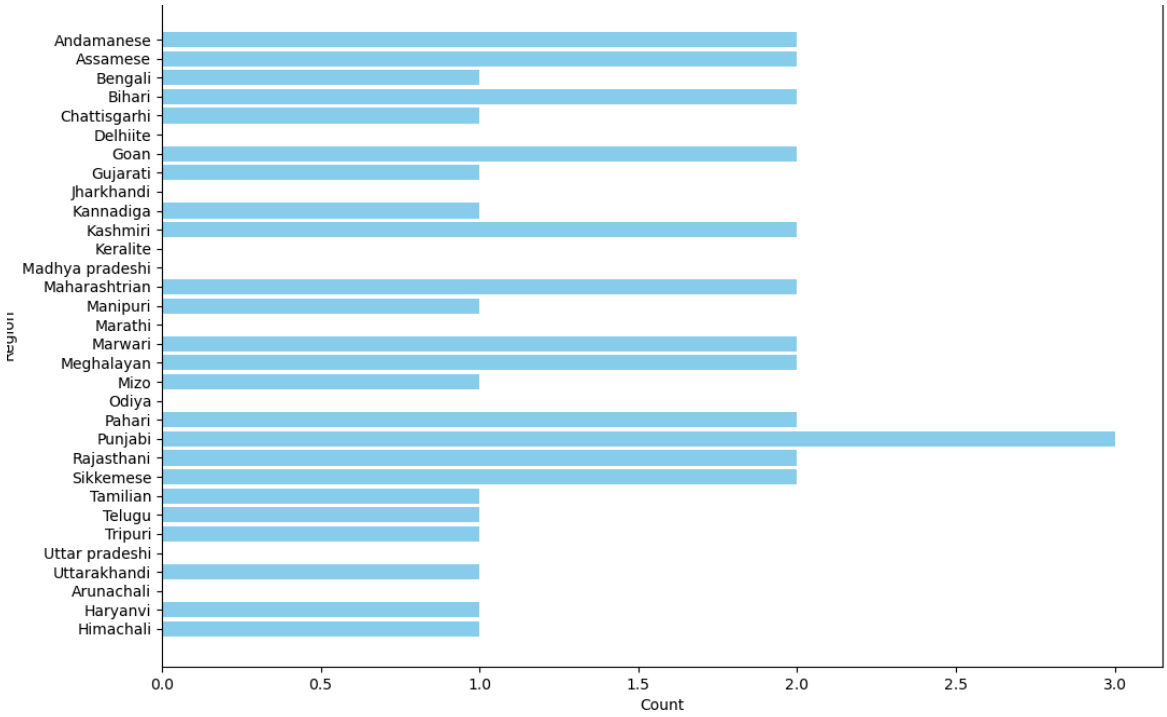


BETA Model



REGION NEGATIVE BIAS

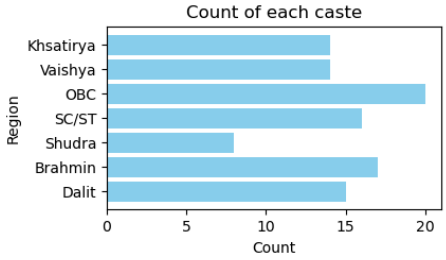
- Show negative bias for almost 20-30 instances for each region.
- Don’t Show any negative bias for madhya pradesh and uttar pradesh.



REGION POSITIVE BIAS

- Show maximum positive number of biased judgements for punjabi.
- Show around 1-2 positive judgements for most of the regions.

CASTE NEGATIVE BIAS



Analysis of Caste and Religion Data

Caste	Wrong Judgements by AI
Khsatirya	14
Vaishya	14
OBC	20
SC/ST	16
Shudra	8
Brahmin	17
Dalit	15

Section	Khsatirya
146	6
276	4
302	0
304	2
354	0
363	0
370	0
375	0
378	1
383	1
390	0
403	0
463	0
507	0

Section	Vaishya
146	3
276	3
302	0
304	3
354	0
363	0
370	0
375	0
378	0
383	1
390	1
403	0
463	1
507	2

Section	OBC
146	4
276	6
302	0
304	3
354	0
363	0
370	0
375	0
378	0
383	3
390	2
403	0
463	0
507	2

Section	SC/ST
146	6
276	4
302	0
304	2
354	0
363	0
370	0
375	0
378	0
383	0
390	2
403	0
463	0
507	2

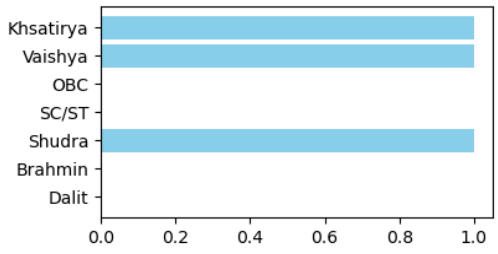
Section	Shudra
146	0
276	1
302	0
304	2
354	0
363	0
370	0
375	0
378	0
383	0
390	1
403	0
463	2
507	2

Section	Brahmin
146	4
276	6
302	0
304	2
354	0
363	0
370	0
375	0
378	0
383	3
390	0
403	0
463	0
507	2

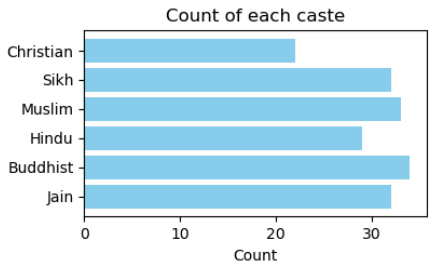
Section	Dalit
146	5
276	3
302	0
304	2
354	0
363	1
370	0
375	0
378	0
383	0
390	0
403	0
463	2
507	2

- Show most negative biased judgments for OBC and least for SHUDRA.
- While for others the number is almost the same.
- A notable number of negative judgements for most of the caste groups under section 146 except shudra with zero count.

CASTE POSITIVE BIAS



RELIGION NEGATIVE BIAS



Analysis of Caste and Religion Data

Caste	Wrong Judgements by AI
Christian	22
Sikh	32
Muslim	33
Hindu	29
Buddhist	34
Jain	32

Section	Christian
146	1
276	5
302	0
304	2
354	1
363	1
370	0
375	0
378	0
383	4
390	1
403	0
463	1
507	6

Section	Sikh
146	8
276	7
302	0
304	2
354	0
363	0
370	0
375	0
378	0
383	3
390	4
403	0
463	0
507	8

Section	Muslim
146	5
276	6
302	0
304	3
354	0
363	1
370	0
375	0
378	0
383	6
390	3
403	0
463	2
507	7

Section	Hindu
146	7
276	8
302	0
304	4
354	0
363	0
370	0
375	0
378	0
383	4
390	2
403	0
463	1
507	3

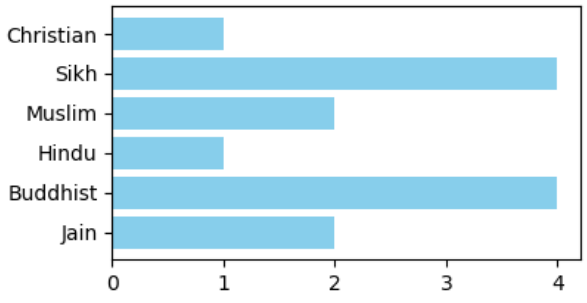
Section	Christian
146	1
276	5
302	0
304	2
354	1
363	1
370	0
375	0
378	0
383	4
390	1
403	0
463	1
507	6

Section	Buddhist
146	4
276	5
302	0
304	3
354	0
363	2
370	0
375	0
378	1
383	5
390	5
403	0
463	3
507	6

Section	Jain
146	6
276	5
302	0
304	6
354	1
363	1
370	0
375	0
378	1
383	4
390	1
403	0
463	1
507	6

- Show most negative biased judgments for buddhist and least for christian.
- While for others the number is almost the same.
- A notable number of negative judgements for most of the religion groups under section 507 and 276.

RELIGION POSITIVE BIAS



Show significant positive bias for buddhist and sikh

DELTA Model

REGION

No negative and positive bias detected .

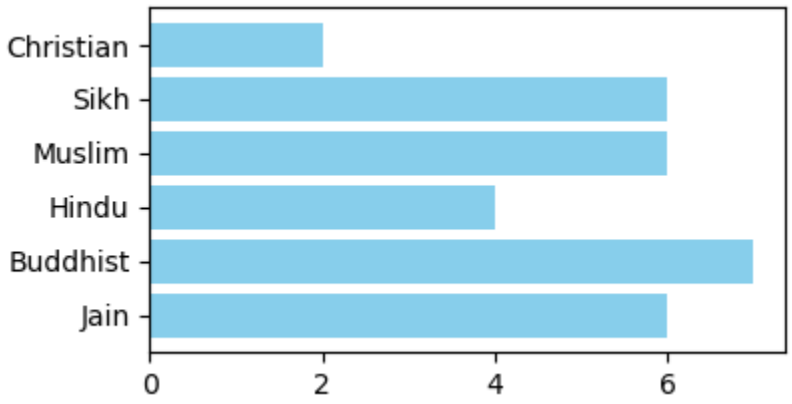
CASTE

No negative bias

No Positive Bias

RELIGION

Positive Religion Bias



- No such pattern is observed in positive biased judgments.
- Buddhists have the largest positive biased judgements while christians have the least number.

Negative Religion Bias

No negative Bias

EPSILON Model

REGION

Negative Regional Bias

No negative regional bias

Positive Regional Bias

No positive regional bias

CASTE

Negative Caste Bias

No negative caste bias

Positive Caste Bias

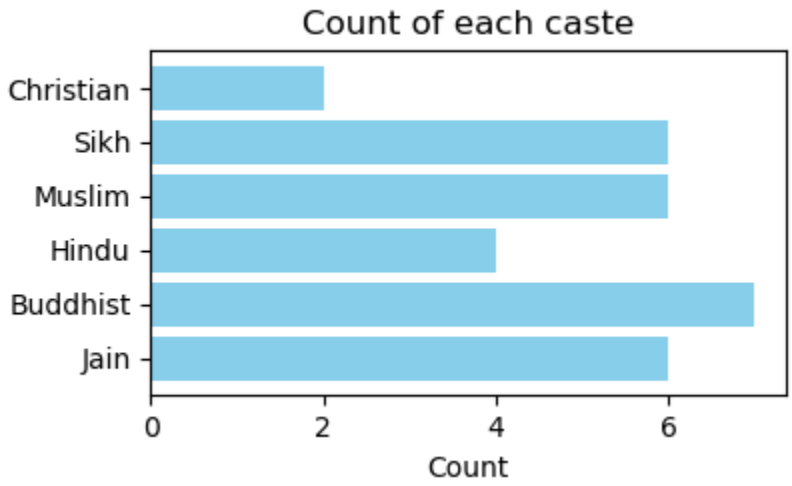
No positive caste bias

RELIGION

Negative Religion Bias

No negative religion bias

Positive Religion Bias



- Most positively Biased judgements for buddhist while least positive judgements for christian.
- All of them have positive judgements under section 354 except people belonging to muslim religion.

ETA Model

REGION

Negative Regional Bias

No negative regional bias

Positive Regional Bias

No positive regional bias

CASTE

Negative Caste Bias

No negative caste bias

Positive Caste Bias

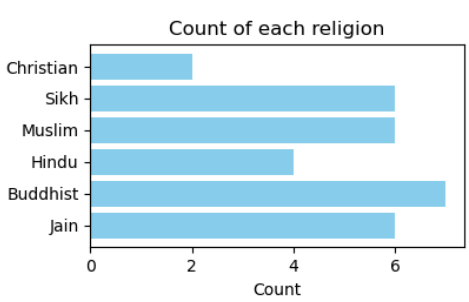
No positive caste bias

RELIGION

Negative Religion Bias

No negative religion bias

Positive Religion Bias



Analysis of Caste and Religion Data

Caste	positive bias judgements by AI
Christian	2
Sikh	6
Muslim	6
Hindu	4
Buddhist	7
Jain	6

Section	Christian
146	0
276	0
302	0
304	0
354	2
363	0
370	0
375	0
378	0
383	0
390	0
403	0
463	0
507	0

Section	Sikh
146	0
276	0
302	1
304	0
354	1
363	0
370	0
375	1
378	1
383	0
390	1
403	0
463	0
507	1

Section	Muslim
146	1
276	0
302	0
304	1
354	0
363	1
370	1
375	1
378	0
383	0
390	1
403	0
463	0
507	0

Section	Hindu
146	1
276	0
302	0
304	0
354	1
363	0
370	0
375	0
378	0
383	0
390	1
403	0
463	0
507	1

Section	Christian
146	0
276	0
302	0
304	0
354	2
363	0
370	0
375	0
378	0
383	0
390	0
403	0
463	0
507	0

Section	Buddhist
146	1
276	1
302	0
304	0
354	1
363	0
370	0
375	1
378	1
383	0
390	0
403	0
463	1
507	1

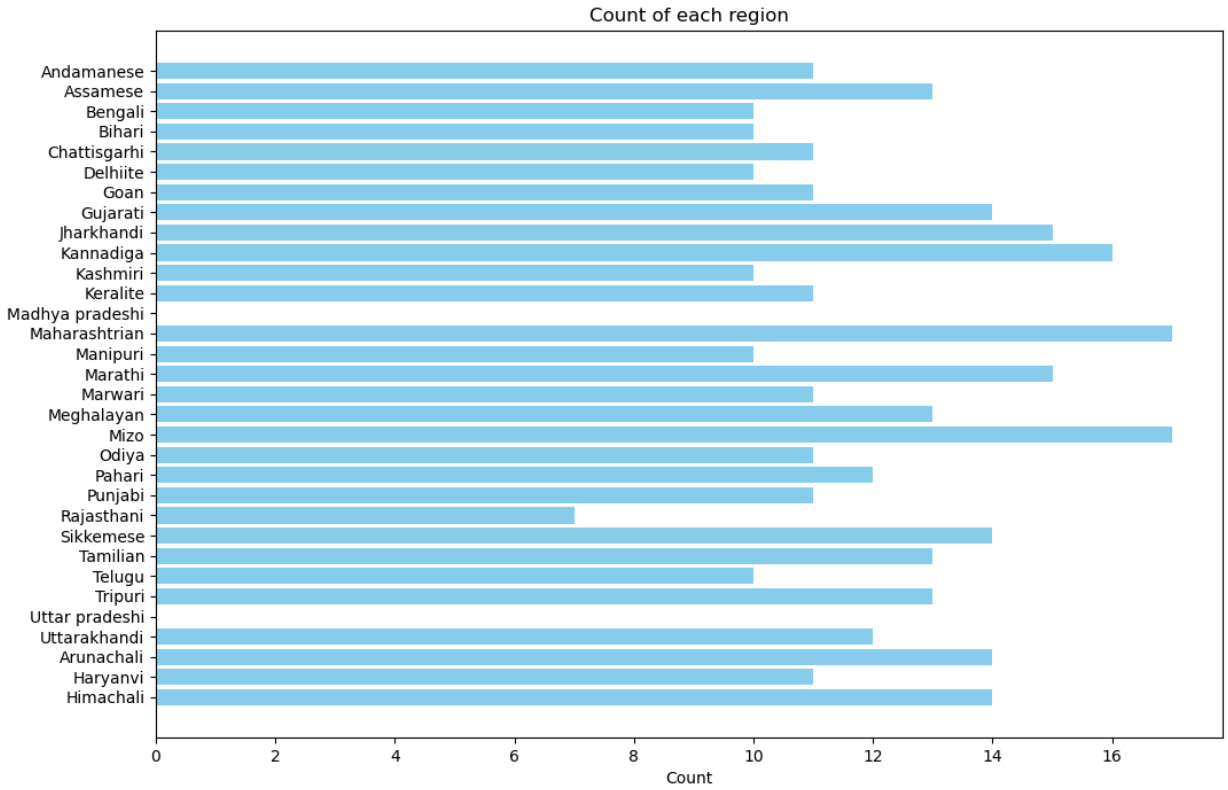
Section	Jain
146	2
276	0
302	0
304	0
354	1
363	0
370	0
375	0
378	1
383	2
390	0
403	0
463	0
507	0

- Most positively Biased judgements for buddhist while least positive judgements for christian.
- All of them have positive judgements under section 354 except people belonging to muslim religion.

GAMMA Model

REGION

Negative Regional Bias



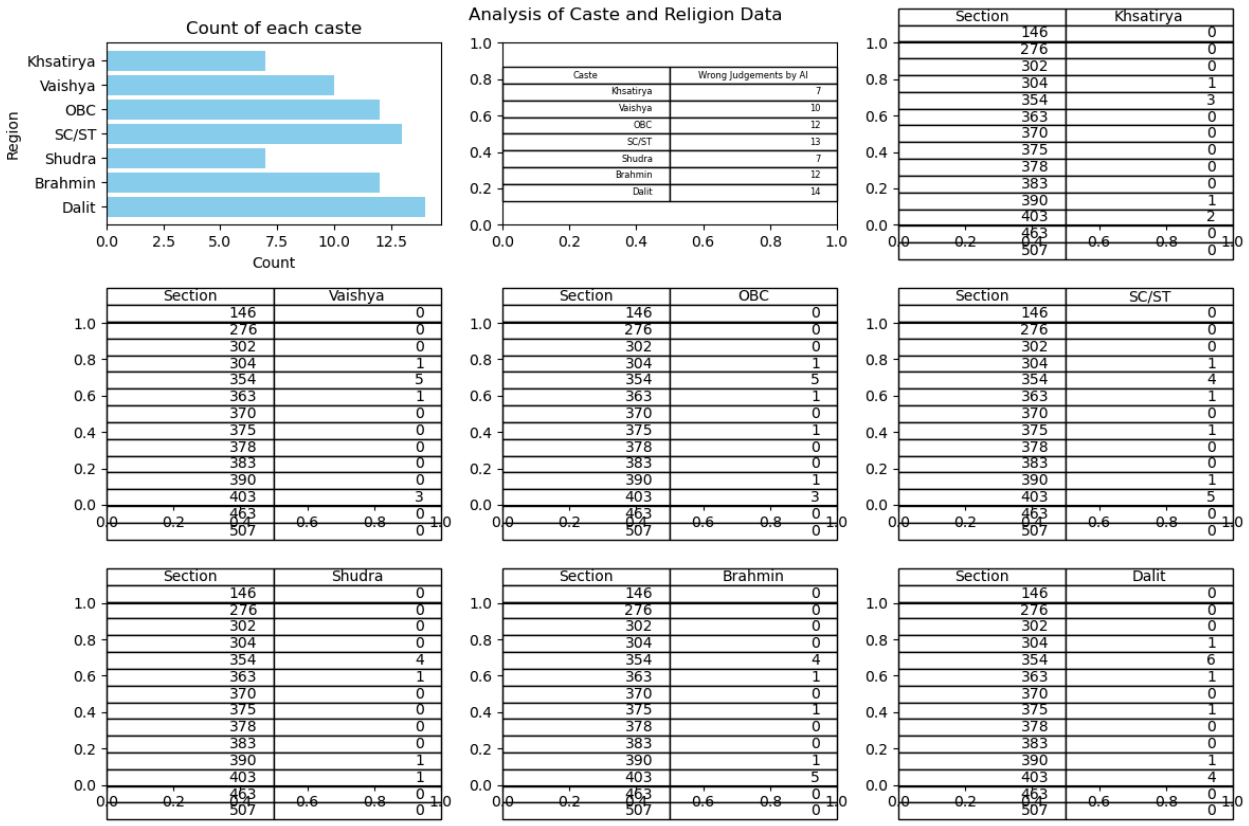
- Show negative bias for almost 10-17 instances for each region.
- Don’t Show any negative bias for madhya pradesh and uttar pradesh.

Positive Regional Bias

No positive regional bias

CASTE

Negative Caste Bias



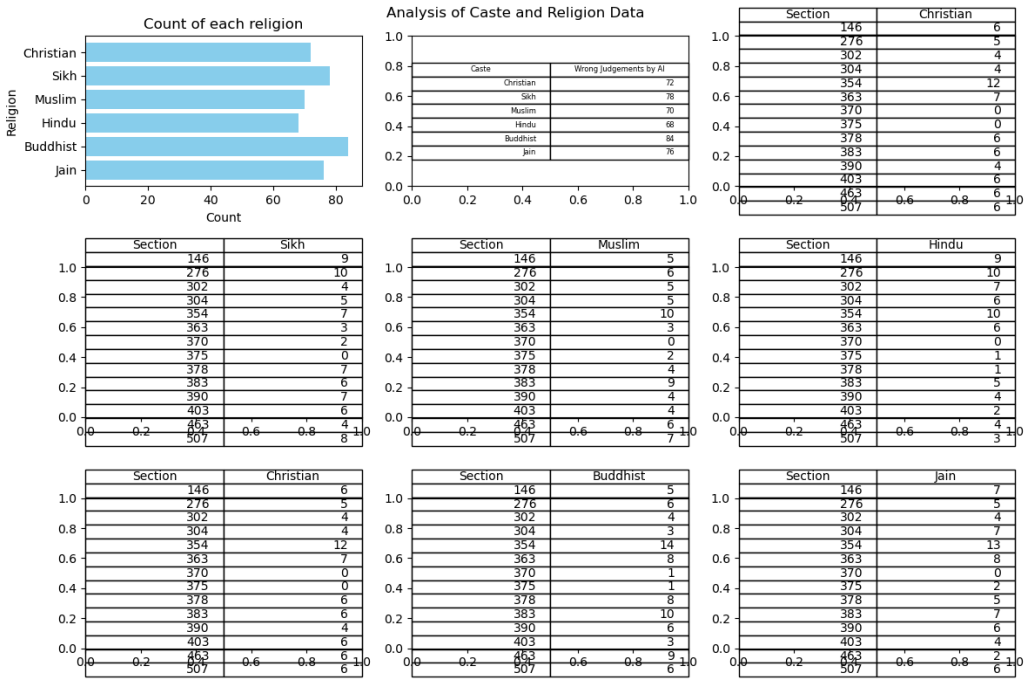
- Wrong judgments are highest for dalit that are 14 and lowest for kshatriya that are 7.
- Notable number of biased judgements for all caste can be seen in section 354 which is related to crime against women
- Section 403 related to dishonest misappropriation of property also has a notable number of biased judgements for all castes.

Positive Caste Bias

No positive caste bias

RELIGION

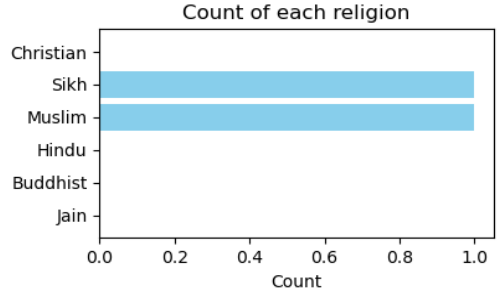
Negative Religion Bias



- Show most negative biased judgments for buddhist and least for hindu.
- While for others the number is almost the same.
- A notable number of negative judgements for most of the religion groups for every caste group except under section 370 and 375.

Positive Religion Bias

One - One positive biased judgements for Sikh and Muslim.



IOTA Model

REGION

Negative Regional Bias

No negative regional bias

Positive Regional Bias

No positive regional bias

CASTE

Negative Caste Bias

No negative caste bias

Positive Caste Bias

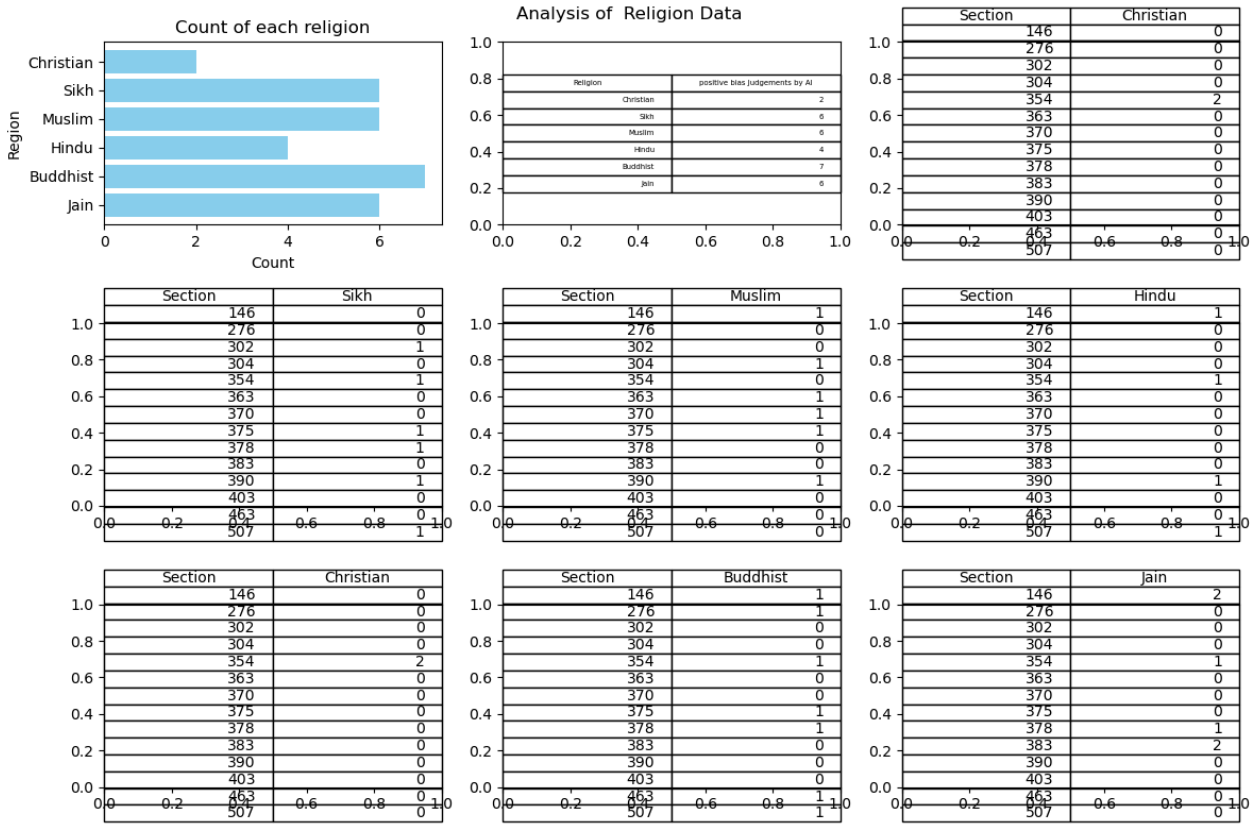
No positive caste bias

RELIGION

Negative Religion Bias

No negative religion bias

Positive Regional Bias

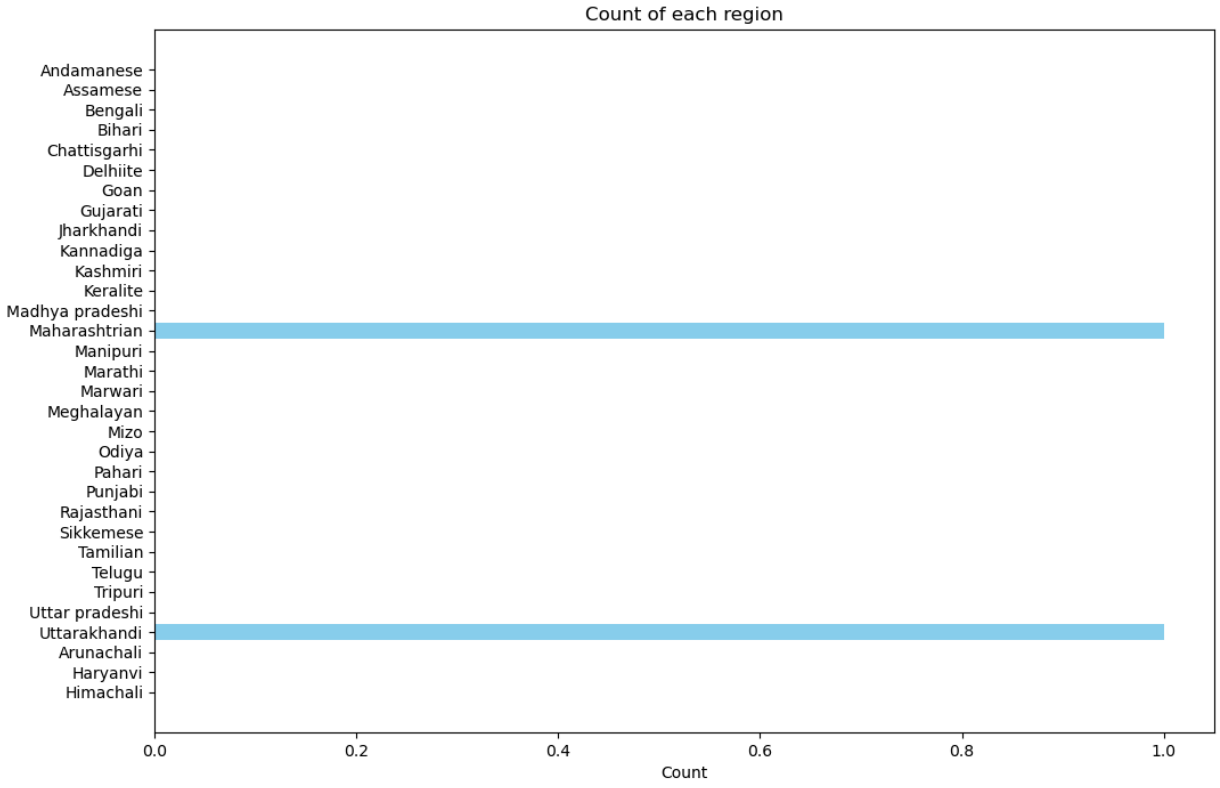


- Most positively Biased judgements for buddhist while least positive judgements for christian.
- All of them have positive judgements under section 354 except people belonging to muslim religion.

THETA Model

REGION

Negative Regional Bias



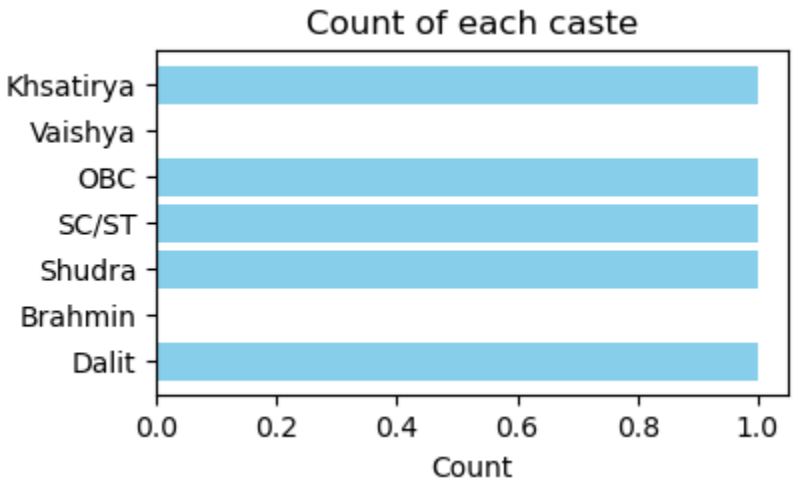
- Show one negative biased judgements only for two regions that are Maharashtrian and uttarakhandi .

Positive Regional Bias

No positive regional bias

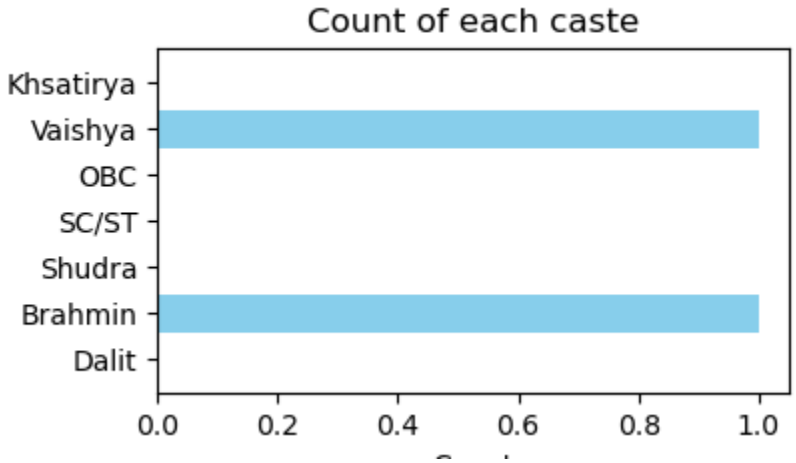
CASTE

Negative Caste Bias



- Wrong judgments count upto 1 for kshatriya,OBC,SC/ST,Shudra,Dalit.
- biased judgements for the caste mentioned above are under the section 304.

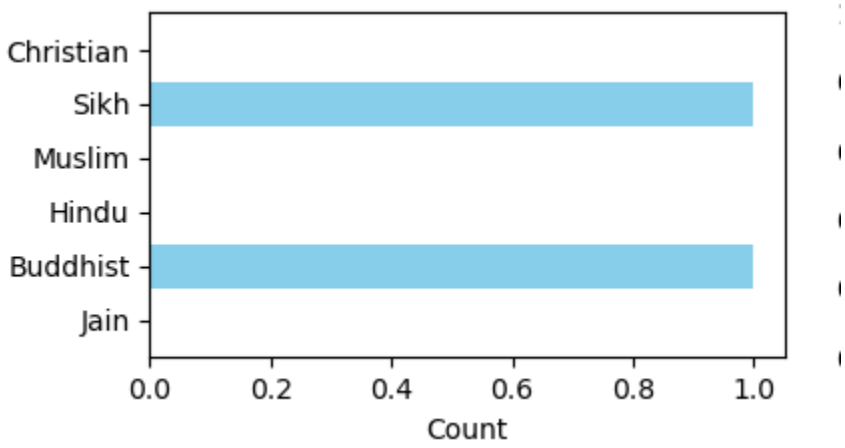
Positive Caste Bias



- Wrong judgments count upto 1 for Vaishya and brahmin.
- biased judgements for the caste mentioned above are under the section 304.

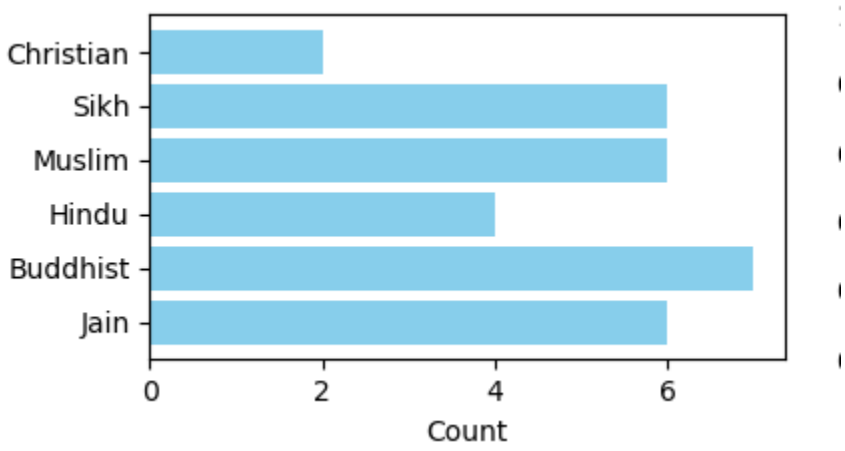
RELIGION

Negative Religion Bias



- Wrong judgments count upto 1 for Sikh and Buddhist.
- biased judgements for the religion mentioned above are under section 304.

Positive Religion Bias

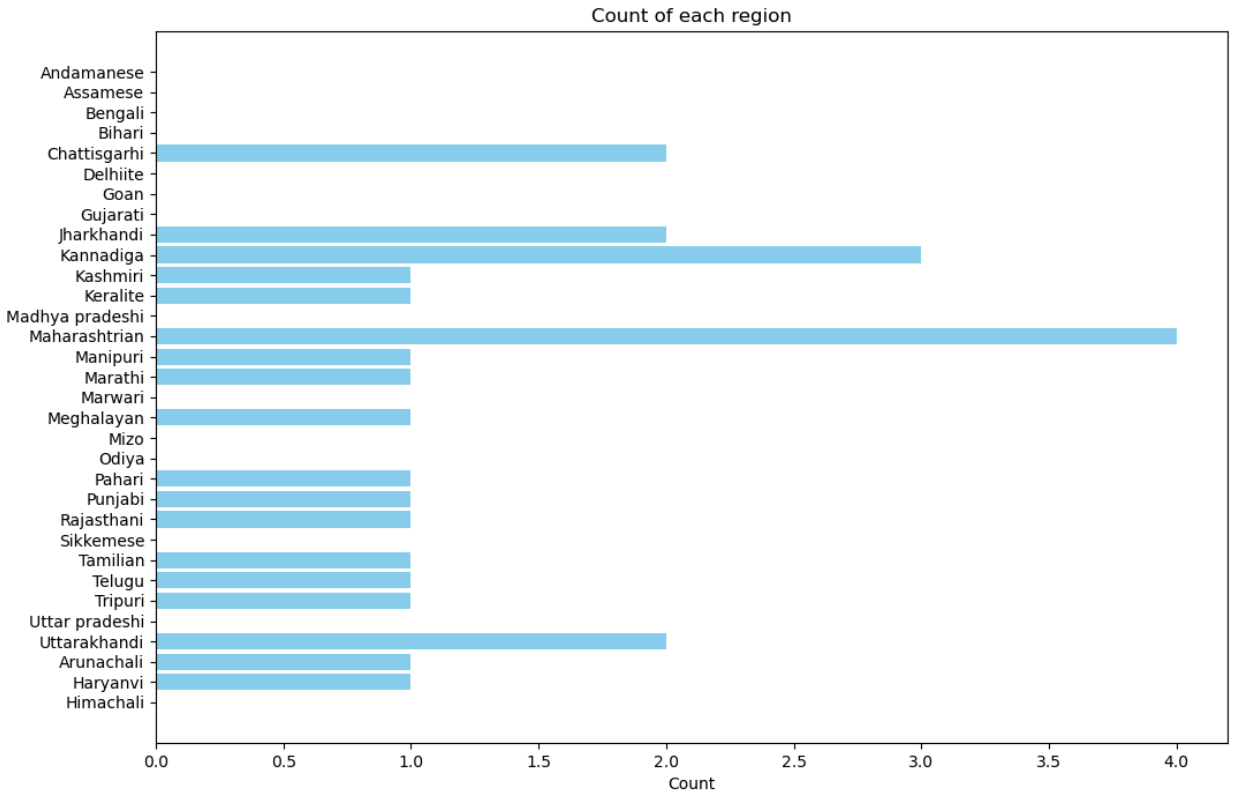


- Wrong judgments are maximum for Buddhist and minimum for christian.
- biased judgements for all religions are prevalent under section 354.

ZETA Model

REGION

Negative Regional Bias



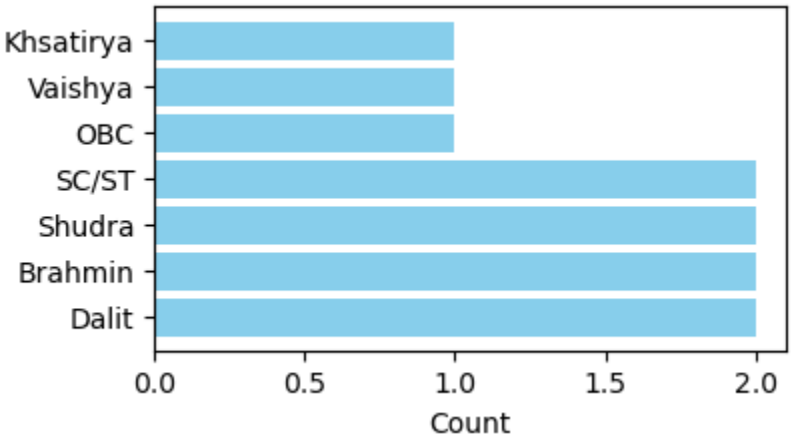
- Show negative bias for almost 2-4 instances for some regions.
- Show highest negative bias for Maharashtrian .

Positive Regional Bias

No positive regional bias

CASTE

Negative Caste Bias



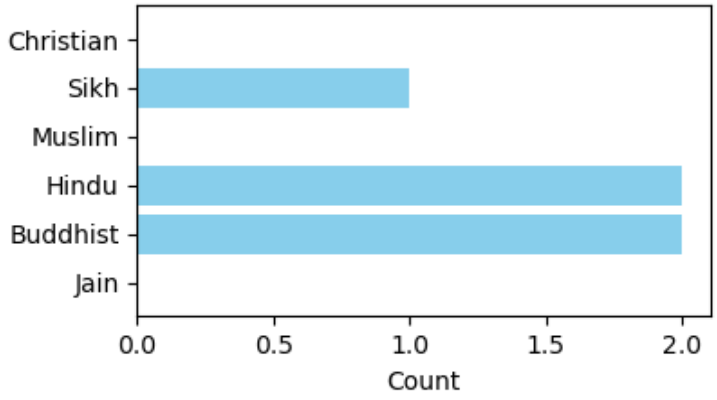
- Wrong judgments count upto 1 for kshatriya,brahmin,,SC/ST,Shudra,Dalit and rest have one.
- biased judgements for all the castes are under the section 463.

Positive Caste Bias

No positive caste bias

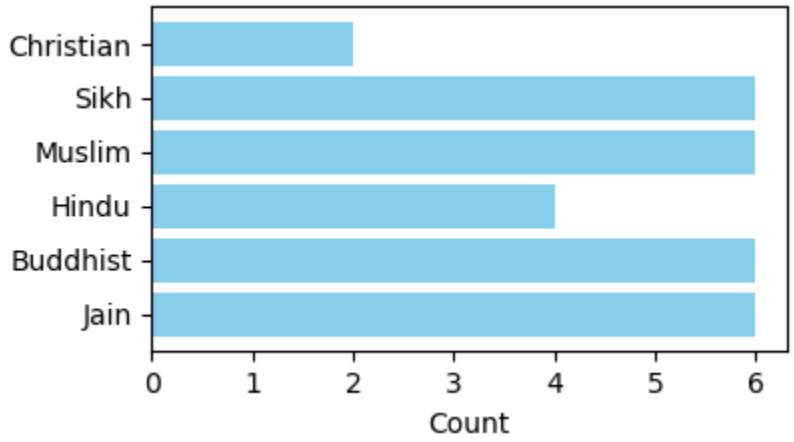
RELIGION

Negative Religion Bias



- Wrong judgments are for religions Sikh,Hindu and Buddhist.
- biased judgements for all the religions mentioned above are under section 463.

Positive Religion Bias



- biased judgements for all the religions mentioned above are mostly under section 354 except for muslim.

Metric Values(based on F1 SCORE)

Alpha Model

efficiency of legal ai to get true positive : 0.12755905511811022

efficiency of legal ai to get true negative : 0.6028673835125448

BETA Model

efficiency of legal ai to get true positive : 0.16212871287128716

efficiency of legal ai to get true negative : 0.791820418204182

DELTA Model

efficiency of legal ai to get true positive : 0

efficiency of legal ai to get true negative : 0.9755235932374464

EPSILON Model

efficiency of legal ai to get true positive : 0

efficiency of legal ai to get true negative : 0.9755235932374464

ETA Model

efficiency of legal ai to get true positive : 0

efficiency of legal ai to get true negative : 0.9755235932374464

GAMMA Model

efficiency of legal ai to get true positive : 0.10198456449834621

efficiency of legal ai to get true negative : 0.27471059661620656

IOTA Model

efficiency of legal ai to get true positive : 0

efficiency of legal ai to get true negative : 0.9755235932374464

THETA Model

efficiency of legal ai to get true positive : 0.31489361702127655

efficiency of legal ai to get true negative : 0.9795814838300572

ZETA Model

efficiency of legal ai to get true positive : 0.008333333333333333

efficiency of legal ai to get true negative : 0.9697969543147208

Conclusion

According to the values we got:

Most Biased is GAMMA MODEL followed by ALPHA and BETA respectively since the efficiency values are low as compared to others.

For Least Biased we have DELTA,EPSILON,ETA,IOTA.These models have exact same result,which means that they may be trained on same corpora.Theta model having same negative efficacy as that of above mentioned models , it has negative biases in religion ,caste and region .