

Does Victim Gender Matter for Justice Delivery? Police and Judicial Responses to Women's Cases in India

Nirvika Jassal

Chuchu Wen & Priscila Stisman

April 3, 2025

- 1 Introduction
- 2 Methods
- 3 Results
- 4 Extensions
- 5 Autopsy
- 6 Suggested Improvements

- **Research Question:** Are women disadvantaged when accessing justice in India compared to men? and if women are discriminated against, is it because of their gender or due to the nature of their complaints?
- **Hypothesis:** Women in India face 'multi-stage' discrimination when accessing justice, encountering systemic barriers at sequential stages of the justice process, including police registration, investigation, trial, and verdict
- **Data:** 418,190 individual-level police reports in Haryana from January 2015 to November 2018 merged with court files

- **OLS Modeling:** to analyze whether women's cases are less likely than men's to be sent to court (+ VAW vs. non-VAW cases)
- **Topic Modeling:** Are there topics in the victims' testimonies that yield low convictions for suspects? Are topics influenced by gender or the classification of the crime as VAW?
- **Topical Inverse Regression Matching**

Structural Topic Models STMs

Why using it?

- Understand the topics in the victims testimonies
- It can help predict whether cases devoted to a topic (e.g. rape) are functions of covariates (e.g. gender)
- Disaggregates what citizens told the police happened to them using statistical associations between words.

Goals

- Give voice to victims by utilizing their own words
- Highlight the severity of claims, especially VAW
- Coarsen high-dimensional data to allow for text matching techniques

FREX (short for Frequency and Exclusivity) is a measure used in STMs to identify words that are both frequent within a topic and exclusive to that topic (highlight words that make a topic more distinguishable from others)

- High Probability Words: These words appear frequently in a topic but might also appear in other topics
- Exclusive Words: These words are more unique to a single topic but might not be the most frequent ones
- FREX Score: FREX balances frequency and exclusivity to identify words that are both common and distinctive for a topic

- **Step 1:** Data Preprocessing. Cleans and tokenizes the text and retains metadata, which contains additional information about each document (e.g., gender of the complainant)
- **Step 2:** Preparing Documents for STM. Refine data, filtering out rare words that appear in fewer than 50 documents
- **Step 3:** Running the STM Model ($K = 32$ topics) in different subsets of the corpus
- **Step 4:** Assign Topic Labels

Run STMs in different corpora and analyze topics within those

- All Crime
- Female Complainant
- VAW Reports

Top Topics Code Example

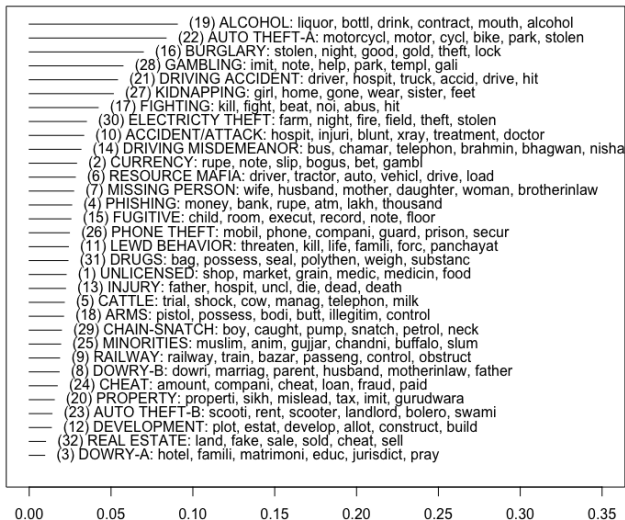
```
set.seed(23456)
processed <- textProcessor(documents = data$text, metadata = data)

out <- prepDocuments(documents = processed$documents,
                     vocab = processed$vocab,
                     meta = processed$meta, lower.thresh = 50)
docs <- out$documents
vocab <- out$vocab
meta <- out$meta
```

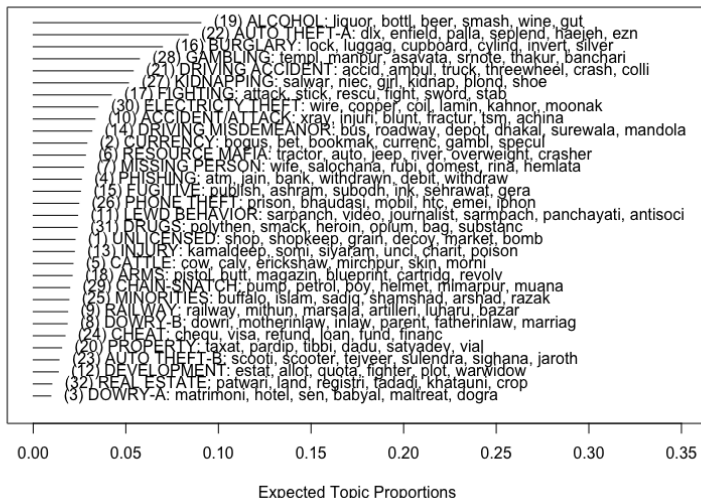
Running Model (1)

```
test2 <- stm(documents = out$documents, vocab = out$vocab,
             K = 32,
             prevalence =~ complainant_gender + gendered + urban + convicted + acquitted + dismissed, # Mult
             max.em.its = 75,
             data = out$meta, init.type = "Spectral",
             verbose = TRUE)
```

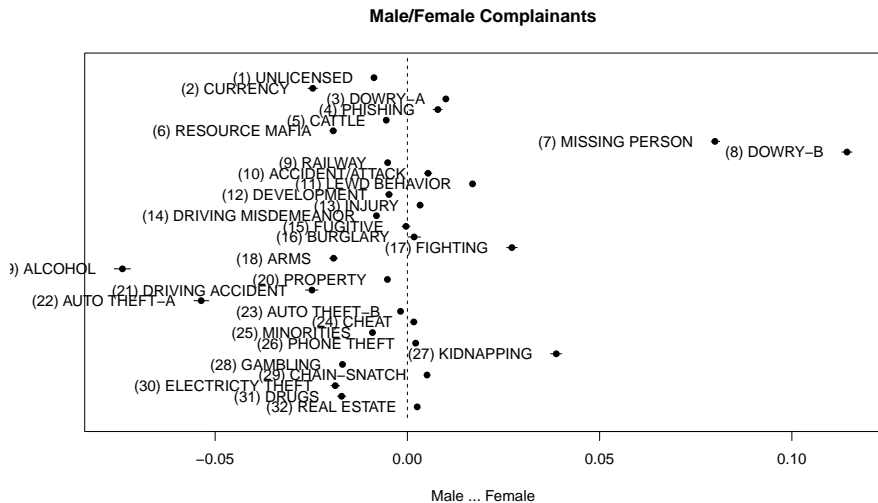
All Crime Top Topics



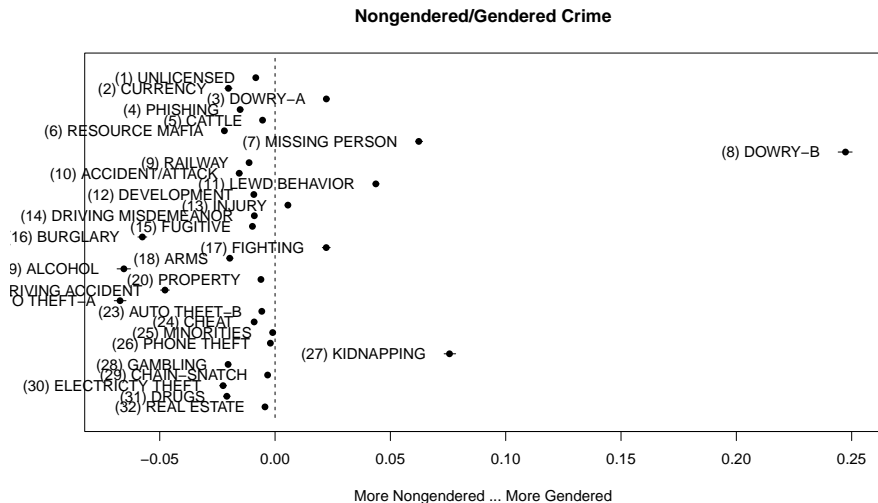
All Crime Top FREX



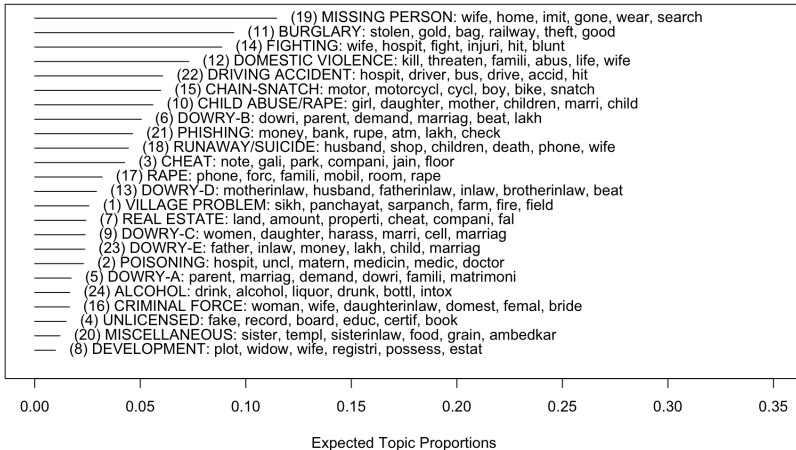
Results



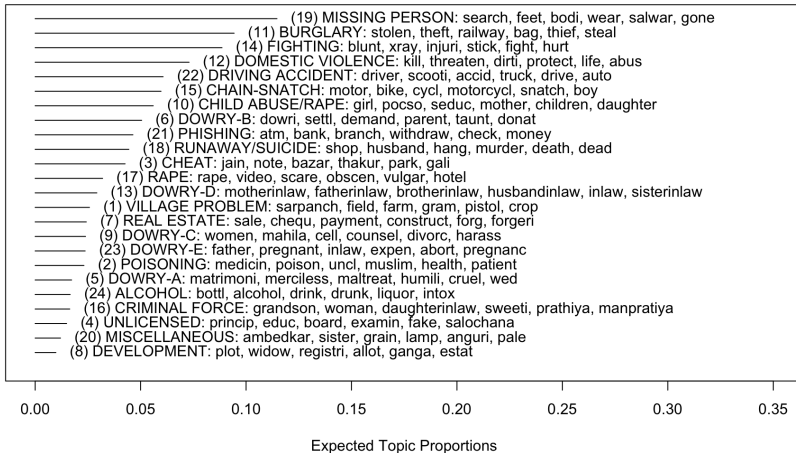
Results



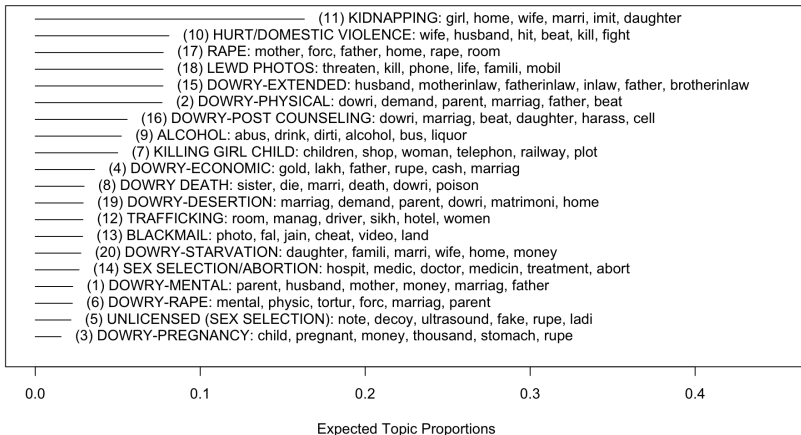
Female Complainant Top Topics



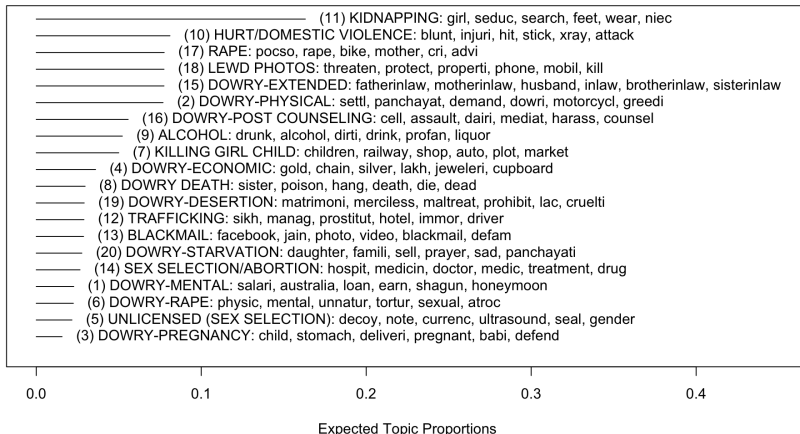
Female Complainant Top FREX



VAW Top Topics



VAW Top FREX



Covariates Code

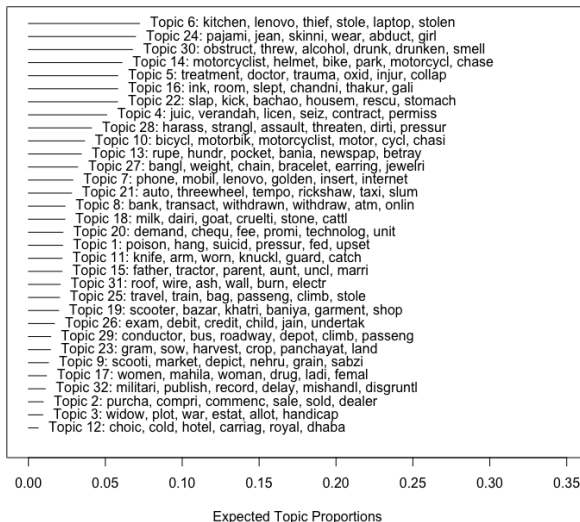
```
set.seed(23456)
processed <- textProcessor(documents = data$text, metadata = data)

out <- prepDocuments(documents = processed$documents,
                     vocab = processed$vocab,
                     meta = processed$meta, lower.thresh = 50)

docs <- out$documents
vocab <- out$vocab
meta <- out$meta

stm.out.c <- stm(
  out$documents,      # The processed documents
  out$vocab,          # The vocabulary of the documents
  K=K,                # Number of topics (32 in this case)
  prevalence=~female_complainant, # Covariate affecting topic prevalence
  content=~female_complainant,    # Covariate affecting topic content
  data=out$meta,          # Metadata associated with the documents
  max.em.its=25,          # Maximum number of EM (Expectation-Maximization) iterations
  seed=1033311           # Random seed for reproducibility
)
```

Non-Gender Crime Content Covariate



- We analyze the most frequently used words by men and women within the same topic in police reports.
- Rationale: Do men and women use certain words differently or with varying frequency when reporting a criminal case?
- For example, in alcohol-related incidents, what words are predominantly used by men, and which ones are more common among women?
- We perform this exercise in three topics in non-gendered police reports

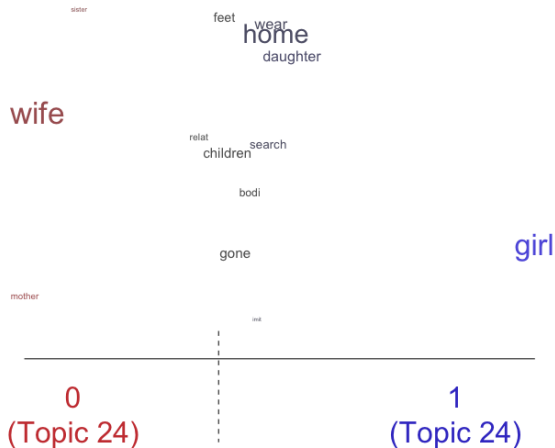
Extensions Results

Word Used for theft related crime, by gender of the complaints



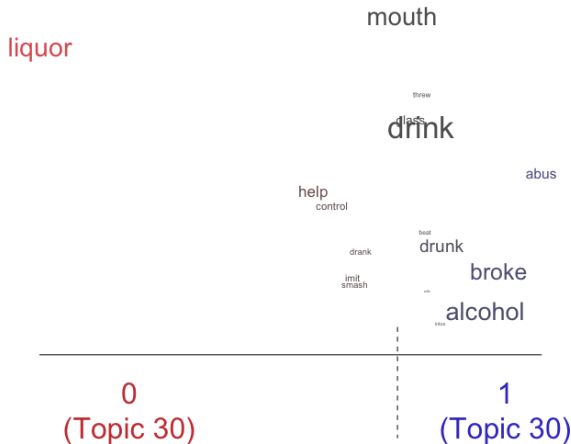
Extensions Results

Word Used for abduction related crime, by gender of the complaints

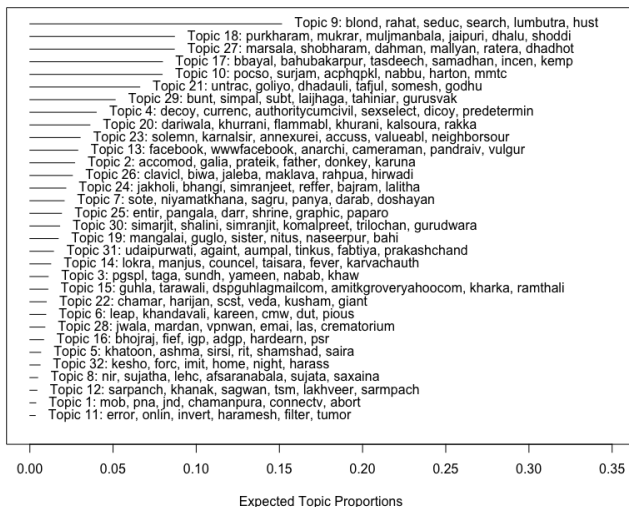


Extensions Results

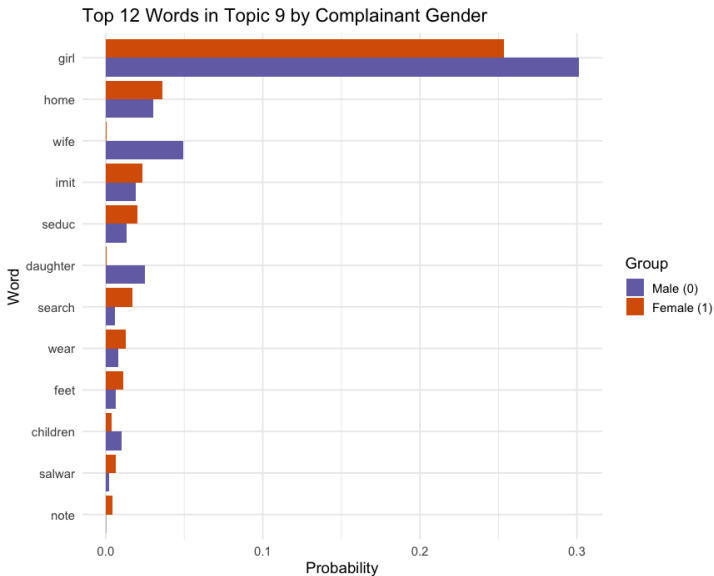
Word Used for alcohol related crime, by gender of the complaints



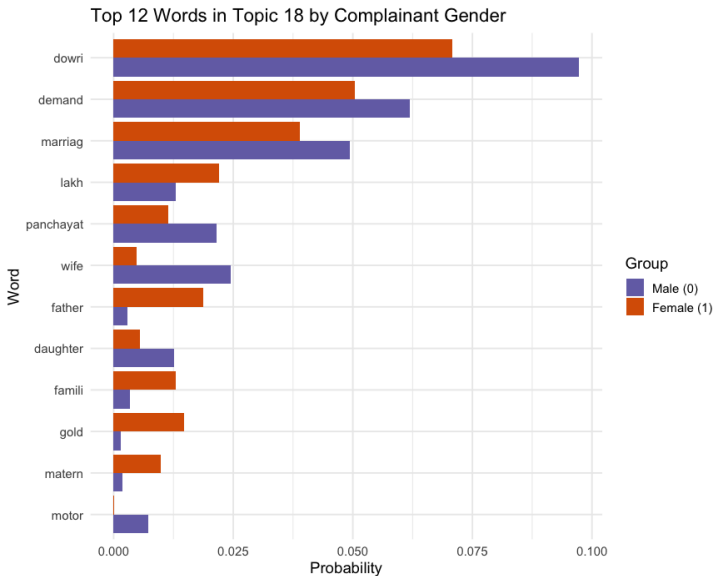
Gendered Crime Content Covariate



Extensions Results



Extensions Results



- Highly satisfied with our STM results
- All STM results were successfully replicated, producing identical outcomes since we used the same seed
- We included the STM for all types of crimes, an analysis presented in the author's code but not in the paper
- Our extensions were straightforward to implement and provided valuable insights
- Great learning experience with STMs, FREX and covariate analysis (metadata, prevalence, content, etc)

Suggested Improvements

- It is unclear how and where the translation from Hindi to English occurs in the author's code
- Part of the corpus is not translated, which makes challenging to easily interpret results
- Lots of typos in the corpus