# Modeling Social Media Narratives about Caste-related News Stories

Prashanth Vijayaraghavan
MIT Media Lab
Cambridge MA, 02139
pralav@media.mit.edu

Lavanya Vijayaraghavan
Fremont, CA 08544
11ya.vijayaraghavan@gmail.com

## Abstract

Caste as a system of social stratification has created a structured culture of oppression in the Indian subcontinent. The rise of social media has paved the way for the democratization of voices providing space for all groups to express, discuss, debate, and form opinions on critical issues. Unfortunately, it also offers a platform for closet or overtly casteist persons to perpetuate discrimination, spread hatred, and sustain casteism under the veil of creating good social media narratives. The overall goal of our work is to model social media narratives associated with caste-specific news stories. To this end, we first aggregate user-generated social media content (e.g., comments) about various caste-related news stories. Next, we analyze these aggregated contents to extract divergent value judgments representing different opinions associated with these news stories. Finally, our ongoing research will provide means to infer the value judgments for each user-generated content automatically, track divergent narratives related to the particular news story, and tackle casteist social media posts using counter-narratives generated by leveraging on the inferred value judgments.



## 1 Introduction

The caste system in the Indian subcontinent is one of the longest surviving social grouping mechanisms characterized by hierarchical gradation of purity. This has led to systematic discrimination and oppression of many communities over several centuries. The Government of India has formally recognized these historically discriminated communities as Scheduled castes, Scheduled Tribes (SC/ST, often referred to as Dalits and Adivasis), and other backward classes (OBCs). A report by the National Dalit Movement for Justice (NDMJ) - National Campaign for Dalit Human Rights, assessed and reported the gaps in the implementation of the law as well as the increase in incidents related to crimes against SC/ST people as recorded by the National Crime Records Bureau from 2009 till 2018. Although social media has opened avenues for civil society, law enforcement agencies, NGOs, people belonging to different social groups to raise awareness about the casteist practices that continue to date, social media also offers space for casteist people to perpetuate their prejudice and demonstrate hostile behavior towards specific sections of the population they consider unequal. Some of these casteist tendencies may arise from privilege, ignorance, implicit bias, radicalized ideologies, or targeted campaigns.



In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Finlayson (eds.): Proceedings of the Text2Story'21 Workshop, Online, 1-April-2021, published at http://ceur-ws.org

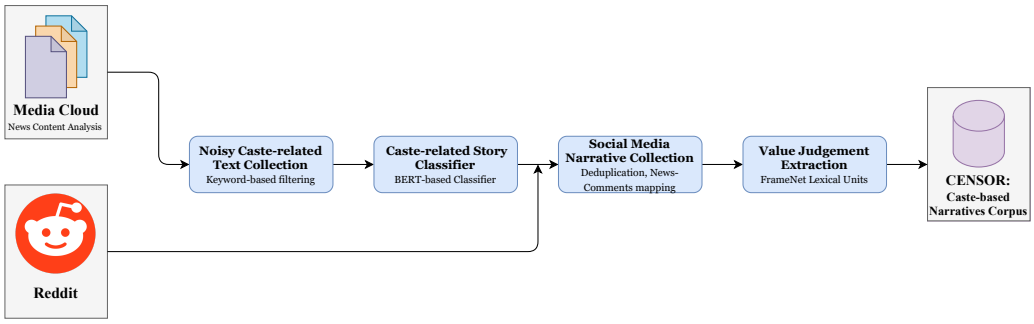


Figure 1: Illustration of our data collection pipeline.

Prior research has primarily focused on studies related to endogamy, and their qualitative interviews and surveys [RMJ19], hate speech detection on social media [KJ18], analysis of caste in newsrooms [FBLM19], to list a few. However, there is a vast gap between its prevalence in our day-to-day life and sufficient NLP efforts to understand and tackle casteism in social media. Given the increasing crimes against the disadvantaged caste sections of the society, this is generally reflected in the news reports despite the bias in newsrooms [FBLM19]. Moreover, several studies have indicated a tremendous growth in online news consumption, especially from social media. However, [WFE+16] observed that exposure to one-sided social media comments with one-sided opinions influenced participants' opinion on issues, mainly when comments contained personal stories. Therefore, we deem it necessary to identify different divergent narratives emerging out of caste-related news stories and tackle casteist discourse with effective counter-narratives.

In our work, we construct a corpus, CENSOR[1], using a data processing pipeline that (a) aggregates caste-related news stories from mainstream news sources, (b) maps these news stories to social media comments (Reddit, in our case), and (b) extracts the value judgements representing different user opinions for the news story. Finally, we will apply different learning strategies to infer value judgments and effectively generate counter-narratives to user-generated comments. Our contribution is three-fold:

- An automated data collection pipeline for aggregating caste-related stories.
- A cross-platform corpus, CENSOR, that contains characteristic narratives aggregated using extraction of value judgments from Reddit content related to caste-related stories.
- Ongoing modeling approach to infer value judgments and generate counter-narratives for user-generated comments.

## 2 Dataset Collection

Figure 1 illustrates our data processing pipeline. The pipeline utilizes the following components to construct the CENSOR corpus:

- **Data Sources**: For our data collection process, we use two data sources – traditional mainstream media and social media. For the former, we refer to a media content analysis platform called Media Cloud[2] that tracks millions of news stories published online and creates an instant analysis of how digital news media covers the topic of our interest. Notably, Media Cloud allows users to choose media sources or collections and submit boolean queries that match these sources' stories. With respect to the social media source, we choose Reddit due to the public availability of data using PushShift API[3].
- **Noisy Caste-related Text Collection**: From traditional news media sources, we specifically lookup for English-language digital news sources and create a keyword-based search for caste. Keywords include broad caste (or varna) names and officially recognized names used by media as explained in Section 1. In this work, we specifically focus on English-language news stories from media collections in India (both national, state, and local level as indicated in Media Cloud). Since the caste atrocities increased during the lockdown

[1]Short for **C**ast**E**-related **N**arrative**S** c**OR**pus
[2]http://mediacloud.org
[3]https://pushshift.io/

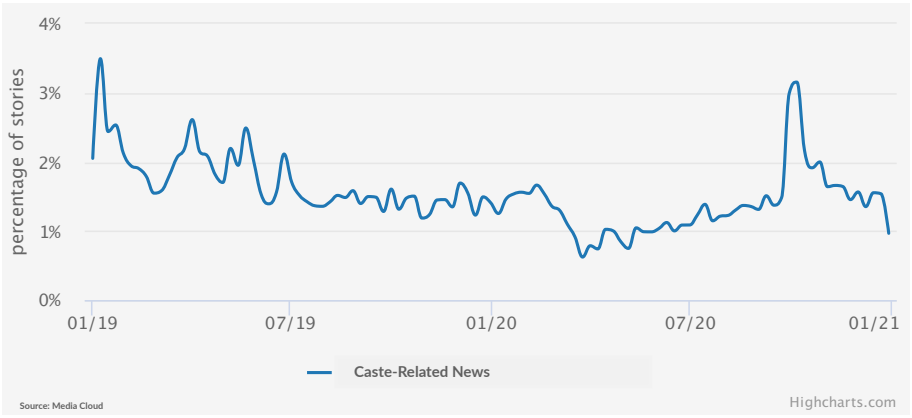


Figure 2: Percentage of caste-related news stories normalized weekly over a 2-year period (Jan'19-Jan'21).

period related to the CoVID-19 pandemic [4], we collect news data over a two-year window, i.e., between 2019 and 2021, to analyze this pattern and characterize the nature of social media narratives during the same period. Given that Media Cloud permits boolean searches, we construct a general boolean query containing keywords related to: (a) general mention of castes or their categories (e.g., castes, Brahmins, Kshatriyas, Vaishyas, Shudras, Dalits, upper castes, lower castes), (b) government recognized group names and their abbreviations (e.g., Scheduled Castes, Scheduled Tribes, SC/ST, Other Backward Classes, OBCs), (c) common references to discriminative practices and forms of oppression (e.g., caste discrimination, untouchability, two-tumbler system, manual scavenging, honour killing, dishonour killing, caste murders, endogamy, inter-caste marriages) and (d) discourse about legal provisions and affirmative action (e.g., reservations, quota, SC/ST act). Figure 2 shows the percentage of caste-related stories normalized weekly over two years from January 2019 to January 2021. We note the peaks refer to important caste-related stories that created a national-level debate. For example, the initial peaks in the first few weeks of Jan'19 are related to EWS quota bill[5] (primarily for the so-called upper castes[6]). We aggregated $\sim 180,400$ news stories using this approach. The aggregated news stories are not restricted to discriminative practices or atrocities but also cover general topics related to castes (e.g., constitutional amendment pertaining to quota bill). Next, we use scrapy[7] to crawl and ingest article content from the URL of these aggregated news stories[8]. Since we used keyword-based search, news stories might not be centered around caste yet picked up by this aggregation method due to the mere occurrence of the term term 'caste'. Hence, this collection could contain other stories that are not centered around castes.

- **Caste-related Story Classifier**: Since we are interested in modeling social media narratives related to caste, the noisy data might draw other irrelevant discussions, which is not our research goal. Therefore, we clean up our data by training a simple binary classifier to identify caste-related news stories. Given that it is a pre-processing step, one of the authors annotated around 1,000 news articles to verify their caste specificity, i.e., tag them being caste-centric or not. Following [DCLT18], we fine-tune a BERT-based model and utilize the representation of the $[CLS]$-token to predict the label probabilities. With an F1 score of 89.6%, we apply this classifier to the noisy collection. This processing results in a total of $\sim 138,800$ news stories.

- **Social Media Narrative Collection**: To collect social media narratives around specific news stories, we use the PushShift API[9] that searches for similar caste-relevant keywords as earlier. We choose only those Reddit posts containing references to URLs from our aggregated news stories dataset. Note that several

[4]https://www.newindianexpress.com/cities/delhi/2020/jul/07/atrocities-against-dalits-see-a-rise-2166477.html

[5]Sample news story– https://economictimes.indiatimes.com/news/politics-and-nation/opposition-questions-timing-of-quota-bill/articleshow/67457766.cms

[6]Usage of terms like "upper" or "lower" relating to caste is meant to describe the existing hierarchy and how the discourse around caste is commonly understood. Such usages are not intended to reinforce that hierarchy.

[7]https://scrapy.org/

[8]We will not make the crawled content data public. However, the URLs and other processed data like topics will be available for future researchers after we wrap up this work.

[9]https://pushshift.io/

Table 1: Samples of relevant expressions extracted from Reddit Comments that contain lexical units provided in their corresponding frames.

| FrameNet Frames | Relevant Reddit Expression |
|---|---|
| Morality_evaluation | It is dishonourable to marry a person from other caste<br>It is good to have an inter-caste marriage |
| Praiseworthiness | It was commendable to stand up against discrimination<br>That's despicable to term everything as an act of discrimination |
| Social_Interaction_Evaluation | It was so cruel that he continued to torture her<br>He was barbaric to torment her when she was vulnerable |

Table 2: Statistics of our aggregated data.

| Dataset Statistics | |
|---|---|
| #Noisy Caste-related Collection | 180,423 |
| #Caste-related Stories | 138,848 |
| #Matched Reddit Posts | 21,589 |
| #Total Comments | 863,560 |
| #Total Comments w/ Value Judgements | 118,776 |

media houses might report a single incident. However, not all of them get equal attention on Reddit. The major incident related to those news stories might get discussed with a single reference to a media story. For those news articles that were shared on Reddit, we use a Python module, PRAW[10], to access Reddit's API and extract user-generated Reddit comments associated with each of these news stories.

- **Value Judgement Extraction**: Towards interpreting different narratives that a news story might lend itself to, it is important to understand varied opinions of right or wrong that users might express based on their beliefs and values. This judgment of the rightness or wrongness of something or someone or the usefulness of something or someone, based on a comparison or other relativity, is referred to as value judgment. A value judgment is a claim about something's moral, practical, or aesthetic worth. E.g., sentences like "That's very righteous of you to help him", "it's unfair to treat someone like that" are normative, which means they evaluate things concerning certain standards or norms. Humans linguistically express such value judgments in many ways. Following [VR21], we explore FrameNet [BFL98] to systematically identify and extract the comments containing expressions of value judgements.Particularly, we chose an abstract frame, referred to as *Social_behavior_evaluation*, involving a speaker judging behavior or action against pre-existing social standards or beliefs for that type of behavior. Other frames inheriting this abstract frame include: *Compliance, Disgraceful_situation, Mental_property, Morality_evaluation, Praiseworthiness Social_interaction_evaluation,* to list a few. Each of these frames contains lexical units that provide various means of expressing value judgments. We extract comments from our Reddit comments to get different value judgments related to the caste-related incident reported in the news story. Table 1 shows samples of relevant expressions extracted from Reddit Comments containing lexical units provided in their corresponding frames. Table 2 shows the data statistics of our aggregated collections.

## 3 Ongoing Modeling Work

To identify different high-level topics associated with caste discourse, we conduct cluster analysis using agglomerative hierarchical clustering using position-weighted Universal Sentence Encoder [CYK+18]. Figure 4 (b) displays the tSNE-visualization of sample stories and cluster labels tagged manually representative of the most common words present in the stories in the cluster. For the news stories in each cluster, we extract different expressions of value judgments from their associated Reddit comments. We aggregate, cluster, and classify them as either

[10]https://praw.readthedocs.io/en/latest/

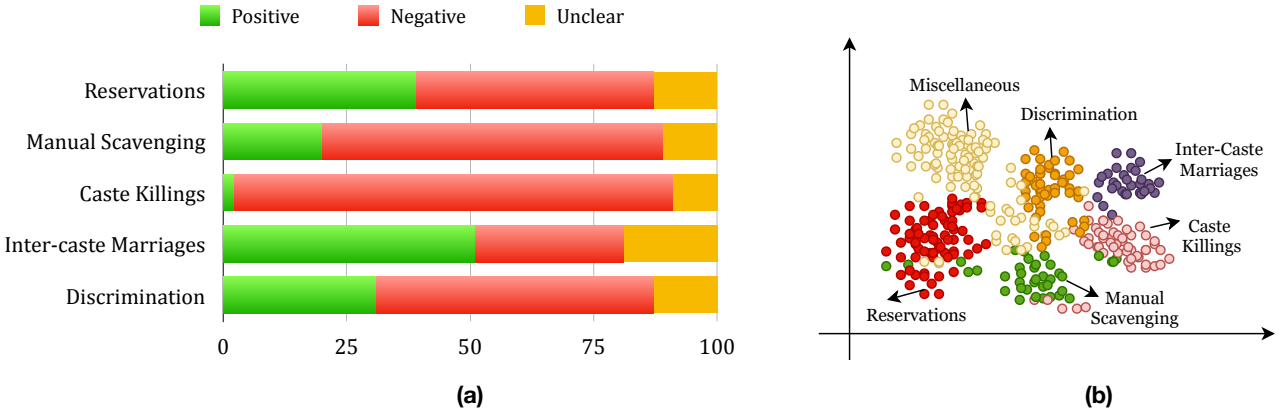


Figure 3: (a) Results of classifying value judgements extracted from Reddit comments associated with each of the news stories in a particular cluster and (b) Visualization of cluster analysis on sampled stories from our dataset.

positive, neutral, or negative at the cluster topic level. We use VADER[11] [HG14] for this purpose and show the results in Figure 4(a). Table 1 also shows the positive and negative value judgment expressions extracted from Reddit comments for some of these categories.

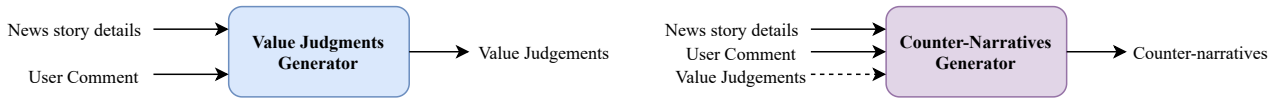


Figure 4: Illustration of the ongoing modeling process.

As a part of our ongoing modeling process (see 4), we investigate ways to identify or generate value judgments by taking news stories and user comments as input. While any language model can be applied to generate value judgments, in this work, we use the transformer language model architecture introduced in [RWC+19] (GPT), which implements multiple transformer blocks of multi-headed scaled dot product attention and fully-connected layers to encode input text [VSP+17]. Next, we will experiment with different controllable generation methods towards generating counter-narratives. We will explore both data and decoding-based methods for this purpose. The former pre-train a language model using the collected dataset, while the decoding-based method introduces a modification to the generation strategy without changes to model parameters. We employ special tokens by prepending them similar to [MSRC20, FG17] or condition on the generated value judgments by concatenation with special delimiter [$SEP$] tokens in between. Our counter-narrative generations can gain from a domain adaptive pretraining strategy (DAPT) [GMS+20] in data-based setting. For the decoding-based method, we will leverage the plug-and-play language modeling work by Dathathri et al. [DML+19], which operates on a pre-trained language model by modifying the current and old hidden state to provide necessary control attribute. Though this process is computation-intensive, it can be more promising than the attribute conditioned method explained earlier. Our evaluation will utilize both automatic and manual generation metrics assessing the relevance of counter-narratives and the quality of generations.

## 4 Conclusion

In this work, we present a method to automatically aggregate caste-related narratives corpus across platforms – mainstream news sources and social media. Further, we intend to model the social media narratives by leveraging expressions of value judgments for generating counter-narratives for caste-related user-generated posts on social media. We plan to run both manual and automatic evaluation metrics on the generated counter-narratives. We believe that our work on modeling discourse around casteism is one of the early efforts to tackle such a critical issue and will springboard further research in this direction.

[11] https://github.com/cjhutto/vaderSentiment

## References


[BFL98] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, 1998.

[CYK+18] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[DCLT18] 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.

[DML+19] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

[FBLM19] António Filipe Fonseca, Sohhom Bandyopadhyay, Jorge Louçã, and Jaison Manjaly. Caste in the news: a computational analysis of indian newspapers. *Social Media+ Society*, 5(4):2056305119896057, 2019.

[FG17] Jessica Ficler and Yoav Goldberg. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[GMS+20] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

[HG14] Clayton J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eytan Adar, Paul Resnick, Munmun De Choudhury, Bernie Hogan, and Alice H. Oh, editors, *ICWSM*. The AAAI Press, 2014.

[KJ18] Satyajit Kamble and Aditya Joshi. Hate speech detection from code-mixed hindi-english tweets using deep learning models. *arXiv preprint arXiv:1811.05145*, 2018.

[MSRC20] Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. Powertransformer: Unsupervised controllable revision for biased language correction. *arXiv preprint arXiv:2010.13816*, 2020.

[RMJ19] Ashwin Rajadesingan, Ramaswami Mahalingam, and David Jurgens. Smart, responsible, and upper caste only: Measuring caste attitudes through large-scale analysis of matrimonial profiles. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 393–404, 2019.

[RWC+19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[VR21] Prashanth Vijayaraghavan and Deb Roy. Modeling human motives and emotions from personal narratives using external knowledge and entity tracking. In *Proceedings of The Web Conference 2021*, 2021.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[WFE+16] Holly O Witteman, Angela Fagerlin, Nicole Exe, Marie-Eve Trottier, and Brian J Zikmund-Fisher. One-sided social media comments influenced opinions and intentions about home birth: An experimental study. *Health Affairs*, 35(4):726–733, 2016.