A COMPUTER-BASED APPROACH TO ANALYZE SOME ASPECTS OF THE POLITICAL ECONOMY OF POLICY MAKING IN INDIA

ANIRBAN SEN

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI

FEBRUARY 2021

A COMPUTER-BASED APPROACH TO ANALYZE SOME ASPECTS OF THE POLITICAL ECONOMY OF POLICY MAKING IN INDIA

by

ANIRBAN SEN

Department of Computer Science and Engineering

Submitted

in fulfillment of the requirements of the degree of

Doctor of Philosophy

to the

Indian Institute of Technology Delhi

FEBRUARY 2021

Certificate

This is to certify that the thesis titled **A Computer-based Approach to Analyze Some Aspects of the Political Economy of Policy Making in India** being submitted by **Mr. Anirban Sen** for the award of **Doctor of Philosophy** in **Computer Science and Engineering** is a record of bona fide work carried out by him under my guidance and supervision at the Department of Computer Science and Engineering, Indian Institute of Technology Delhi. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

Dr. Aaditeshwar Seth
Associate Professor
Department of Computer Science and Engineering
Indian Institute of Technology Delhi
New Delhi- 110016

Acknowledgments

Growing up in a middle class family, a meaningful life looked pretty straightforward to me. Get a degree, get a job, start a family and you would have accomplished enough. I happened to disagree. Eight years later, after quitting my job to get back to academics, here I am, submitting my PhD thesis, and I am not sure if I will be able to do justice to this section as the support and help I have received throughout is uncountable and immense.

First of all, I thank God for giving the capability to pursue my research and for blessing me with my parents and my aunt, without whose motivation, I would not even have dreamt of embarking on this journey. I left home to pursue my dream, and never did they hold me back at any point. I have missed birthdays, trips and even doctor's appointments, and they still have continued to support me without a question. My gratitude will never be enough for the family I have.

Words fail me to describe the kind of guidance I have received from my supervisor, Dr. Aaditeshwar Seth who has not only guided me through my academic research, but has pushed me to evolve into a more idealistic and sensitive human being. It is an understatement to say that I look up to him as a researcher and an academician, but as a teacher, a mentor, and an individual dedicated towards contributing to the society. I am also grateful to my SRC members, Dr. Reetika Khera, Dr. Amitabha Bagchi and Dr. Rahul Garg for their valuable advice and feedback. I have been fortunate to work with multiple brilliant members of the ACT4D team at IIT Delhi, without whose participation, my work would not have taken the shape it has.

I take this opportunity to also thank my collaborators at Ashoka University, Dr. Priyamvada Trivedi and Saloni Bhogale, for their useful insights and data assistance for my research. A good part of my development as a researcher happened during my internship days at Xerox Research Centre India, and I will remain forever grateful to Dr. Shourja Roy, Dr. Sandya Mannarswamy and Dr. Manjira Sinha for their support and guidance at that time. In this life changing journey of research, I have travelled to two different continents other than Asia, and the experiences that I have gathered have been nothing less than inspiring. I would like to thank the TCS Research Scholar Program, ACM IARCS, and Microsoft Research for all the travel grants and enthusiasm.

To be honest, it has not always been perfect, but quite the opposite. I would have probably not made this far if it were not for my friends Dipanjan, Omkar, Prajna, Madhulika, Debjyoti, Santanu, Siddhartha, and Indrabati. The last one, Indrabati, being my partner, has had to spend sleepless nights with me worrying over deadlines and work pressure; and has even put up with me being an absent spouse at times. I feel blessed for finding yet another family in her mother who has been a source of encouragement all along.

Ideas are like seeds which once sown in your mind, shall grow to become a tree. The seed of research was sown in my mind by my master's guide, Dr. Saptarshi Ghosh.

Finally, I have to admit that it was a dream come true to have been able to be a part of the IIT Delhi fraternity. I thank each and every member of the Computer Science and Engineering Department and the administrative staff of IIT Delhi (specifically Arun Sir, Rajesh Sir, and Suresh Sir from the School of IT). This was an experience of a lifetime and I could not have been more thankful for all the opportunities and the long walks in the campus. I might have missed to name many but I remain grateful for every advice, every word of encouragement, and every blessing that has kept me going. I feel blessed.

Anirban Sen

Abstract

To understand the policy process appropriately, it is essential to obtain a bird's eye view of the political economy around policies. According to Wikipedia, 'Political economy is the study of production and trade and their relations with law, custom and government; and with the distribution of national income and wealth'. Analysis of political economy deals with the study of relationships between the government and the market, which includes corporations, business-persons, and other corporate entities related to trade and production. We develop a technological platform to analyze some facets of the political economy around important policies in India, which in turn shall aid citizens in obtaining a good understanding of the policy issues. The contribution of this thesis is that it shows that such analysis can be performed using publicly available web and media data, using a suite of computer scientific tools and techniques.

There are different facets of political economy analysis. In this thesis, we study a few of them, namely the interlocks between the corporate and state entities, the kind of statements made by these entities in popular media, the bias in policy representation carried by the mass media and the social media, and the policy discourse that occurs in

these media sources and in the Parliament. Our findings suggest that interlocks between corporate and government entities are increasing over time, which could lead to their increasing influence in policy-making. We also find that the mass media is biased towards various entities and topics relevant to the policies, and that the representation of policies in the mass media and the Parliament does not equitably cover issues of all sections of people. Moreover, policy discourse in popular media chiefly includes the views of politicians and business-persons, and does not provide adequate attention to the views of policy experts or academicians who can provide valuable insights on technical nuances and problems in policy implementation. Social media is seen to accentuate the biases carried by mass media, and it closely follows the topics covered by the mass media regarding various policies. Additionally, we also propose a novel news recommendation algorithm in this thesis, which can counter the issue of algorithmic bias by ensuring fairness and diversity in representation of various topics relevant to the policies.

While several studies have already been done in the domain of political economy analysis, this thesis is the first work that attempts to analyze some aspects of the political economy around policy-making, in the Indian context, using computer scientific techniques on large scale, and publicly available data. Our findings from this work have been updated in a website with the objective of reaching a target audience of journalists, social activists, policymakers, researchers and citizens in general. We believe that the technological platform suggested in this thesis can serve to make people more aware of the political economy that affects policy-making, peoples' opinion on these policies, the democracy, and ultimately their lives and lives of others.

सारांश

नीति प्रक्रिया (Policy Process) को उचित रूप से समझने के लिए, नीतियों से संबंधित राजनीतिक अर्थव्यवस्था (Political Economy) के बारे में विहंगम दृष्टि प्राप्त करना आवश्यक है। विकिपीडिया के अनुसार , `राजनीतिक अर्थव्यवस्था उत्पादन और व्यापार, तथा कानून, प्रथा, सरकार और राष्ट्रीय आय और धन के वितरण के साथ उनके संबंधों का अध्ययन है'। राजनीतिक अर्थव्यवस्था का विश्लेषण सरकार और बाजार के बीच संबंधों के अध्ययन से संबंधित है, जिसमें कंपनी, व्यवसायी तथा व्यापार और उत्पादन से संबंधित अन्य कॉर्पोरेट इकाइयां शामिल हैं। हमने भारत में महत्वपूर्ण नीतियों से संबंधित राजनीतिक अर्थव्यवस्था के कुछ पहलुओं का विश्लेषण करने के लिए एक तकनीकी मंच विकसित किया है, जो नीतिगत मुद्दों की अच्छी समझ प्राप्त करने में नागरिकों की सहायता करेगा। इस थीसिस का योगदान यह है, कि यह दर्शाता है कि इस तरह का विश्लेषण, सार्वजनिक रूप से उपलब्ध वेब और मीडिया डेटा का उपयोग करके, कंप्यूटर वैज्ञानिक उपकरणों और तकनीकों के एक सूट का उपयोग करके किया जा सकता है।

राजनीतिक अर्थव्यवस्था विश्लेषण के विभिन्न पहलू हैं। इस थीसिस में हम उनमें से कुछ का अध्ययन करते हैं, जैसे कॉरपोरेट और राष्ट्रीय संस्थाओं के बीच अंतर सम्बन्ध, लोकप्रिय मीडिया (समाचार पत्र) में इन संस्थाओं द्वारा दिए गए बयान, समाचार पत्र और सोशल मीडिया द्वारा नीतिगत प्रवचन में किए गए पक्षपात, तथा मीडिया और संसद में किए गए नीतिगत चर्चा। हमारे निष्कर्ष बताते हैं कि समय के साथ कॉर्पोरेट और सरकारी संस्थाओं के बीच अंतर सम्बन्ध (interlocks) बढ़ रहे हैं, जिससे नीति–निर्माण में उनका प्रभाव बढ़ सकता है। हम यह भी पाते हैं कि समाचार पत्र विभिन्न लोगों और नीतियों के लिए प्रासंगिक विषयों के प्रति पक्षपाती है , और जनसंचार माध्यमों और संसद में नीतियों का प्रतिनिधित्व सभी वर्गों के लोगों के मुद्दों को समान रूप से कवर नहीं करता है।

इसके अलावा, लोकप्रिय मीडिया में नीतिगत प्रवचन में मुख्य रूप से राजनेताओं और व्यवसायी व्यक्तियों के विचार शामिल होते हैं, और नीति विशेषज्ञों या शिक्षाविदों के विचारों पर पर्याप्त ध्यान नहीं दिया जाता है , जो तकनीकी बारीकियों और नीति कार्यान्वयन में समस्याओं पर मूल्यवान अंतर्दृष्टि प्रदान कर सकते हैं। सोशल मीडिया बड़े पैमाने पर समाचार पत्रों द्वारा किए गए इस पक्षपात को बढ़ाता है, और यह विभिन्न नीतियों के बारे में समाचार पत्रों द्वारा कवर किए गए विषयों का बारीकी से अनुसरण करता है। इसके अतिरिक्त, हम इस थीसिस में एक समाचार एग्रीगेटर एल्गोरिथ्म का भी प्रस्ताव करते हैं, जो नीतियों के लिए प्रासंगिक विभिन्न विषयों के प्रतिनिधित्व में निष्पक्षता और विविधता सुनिश्चित करके एल्गोरिथम में पक्षपात के मुद्दे का मुकाबला कर सकता है।

हालांकि कई अध्ययन पहले से ही राजनीतिक अर्थव्यवस्था विश्लेषण के क्षेत्र में किए गए हैं, यह थीसिस पहला काम है जो कंप्यूटर के उपयोग से भारतीय संदर्भ में नीति–निर्माण के आसपास की राजनीतिक अर्थव्यवस्था के कुछ पहलुओं का, बड़े पैमाने पर वैज्ञानिक तकनीक, और सार्वजनिक रूप से उपलब्ध डेटा की मदद से, विश्लेषण करने का प्रयास करता है। इस काम से हमारे निष्कर्ष एक वेबसाइट में पत्रकारों, सामाजिक कार्यकर्ताओं, नीति निर्माताओं, शोधकर्ताओं और नागरिकों के लक्षित दर्शकों तक पहुंचने के उद्देश्य से अपडेट किए गए हैं। हमारा मानना है कि इस थीसिस में सुझाया गया तकनीकी मंच लोगों को राजनीतिक अर्थव्यवस्था के बारे में अधिक जागरूक बनाने का काम कर सकता है, जो इन नीतियों पर, लोकतंत्र पर, और अंततः उनके जीवन और दूसरों के जीवन पर सकारात्मक प्रभाव ला सकता है।

Contents

Certificate **i**

Acknowledgements **iii**

Abstract **v**

List of Figures **xvii**

List of Tables **xxv**

1 Introduction **1**

1.1 Research Questions 3

1.2 System Architecture 5

1.3 Thesis Structure 7

2 Research Methodology **11**

2.1 Technological System 12

2.1.1 Data Collection 13

2.1.2 Entity Resolution 17

2.1.3 Aspect extraction using LDA 20

2.1.4 Sentiment Analysis 22

2.1.5 Computation of Entity Scores 24

2.2 Qualitative Analysis of Data 24

2.3 Data Presentation 27

3 Related Work **29**

3.1 Political Economy Analysis and Its Importance 29

3.1.1 Foundational Studies on Political Economy Analysis 29

3.1.2 Some Applications of Political Economy Analysis 31

3.2 Frameworks for Political Economy Analysis 33

3.3 Methods of Political Economy Analysis 34

3.4 Media's Impact on Political Behavior 39

3.4.1 Mass Media's Impact on Political Behavior 39

3.4.2 Web and Social Media's Impact on Political Behavior 40

3.5 Analysis of Bias 42

3.5.1 Mass Media Bias 42

3.5.2 Web and Social Media Bias 45

3.6 Representation of Policies in the Parliament 48

3.7 Critical Perspectives of Big Data Analysis 49

4 Analysis of Corporate-Government Interlocks 51

4.1 Related Work 52

4.2 Details of Network Computation 55

4.3 Indicator Monitor Application 56

4.4 Methodological Analysis 58

4.4.1 Special cases of entities captured by our ranking 58

4.4.2 Normalization of node scores 60

4.4.3 Comparing the indicator with a random baseline 61

4.4.4 Randomizing only bridges: minmax normalization 62

4.4.5 Randomizing only bridges: rank normalization 63

4.4.6 Data limitations 65

4.5 What are the causes behind increase in the interlocks? 66

4.6 Discussion and Conclusion 69

5 Analysis of Bias in Mass Media Content 73

5.1 Related Work 74

5.2 Aspect Coverage Bias of Mass Media 81

5.3 Constituency Coverage Bias of Mass Media 85

5.4 Political Party Bias of Mass Media 88

5.5 Alignment With Social Media Content 89

5.6 Discussion and Conclusion 93

6 Analysis of Discourse on Economic Policies 97

6.1 Related Work 98

6.2 Aspects Covered by the Media and the Parliament 99

6.3 Variation in Questions Asked by Political Parties 104

6.4 Discussion and Conclusion 106

7 Analysis of Policy Representation in Mass Media 109

7.1 Related Work 110

7.2 Most Vocal Entities and Groups in Mass Media 116

7.3 Sentiment Slant of Statements by Elites 120

7.4 Top Aspects Covered by Mass Media 124

7.5 Discussion and Conclusion 126

8 Towards a Fairness and Diversity Guaranteeing News Aggregator 129

8.1 Related Work 131

8.2 Data 133

8.3 System Architecture 135

8.3.1 Aspect Identification for a Temporally Evolving Feed 136

8.3.2 Recommendation Algorithm to Ensure Fairness and Diversity . . . 138

8.4 Results 145

8.5 Discussion and Conclusion 153

9 Conclusion and Discussion **155**

9.1 Primary Challenges 157

9.2 Limitations and Future Work 160

9.2.1 Analysis of Interlocks 160

9.2.2 Policy Representation and Bias Analysis 161

9.2.3 News Recommendation 163

9.2.4 Reaching Out to the Users 164

9.3 Recommendation Towards Accountability of Participants of Democracy . . 164

9.3.1 Existing Bodies for Regulation 165

9.3.2 Need for Citizen Led Accountability 168

Bibliography **171**

Appendices **197**

A Analysis of Bias in Mass Media Content **199**

A.1 Relative Coverage of Aspects 199

A.2 Coding Schema 201

A.3 Aspect to Constituency Alignment Matrices 202

A.4 Coverage of Constituencies by Mass Media 203

A.5 Alignment of news-sources with their readers 206

B Towards a Fairness and Diversity Guaranteeing News Aggregator 217

B.1 Calculating the Positive Percentage 217

B.2 Comparison of Inferencing and Retraining 218

B.3 Calculation of the Fairness Window 219

B.4 Performance based on a Modification of the Algorithm 221

B.5 Filtering Event based Articles from Google Alerts 223

B.6 Performance on Highly Skewed Dataset 224

B.7 Results (continued) 225

List of Publications 229

Biography 231

List of Figures

1.1 Overall architecture of the system: The lowermost layer belongs to data collection where web data is crawled from a multitude of sources; the middle layer contains the algorithms used to clean and analyze the data; the uppermost layer contains the applications that we build upon this analysis. 6

2.1 Pipeline to answer: (RQ-1 [Chapter 4]) How can corporate-government interlocks be identified that may have a potential influence on the policy process? 12

2.2 Pipeline to answer: (RQ-2 [Chapter 5]) Is mass media biased in how it represents policies? (RQ-3 [Chapter 6]) Is the policy-making process democratic, i.e., one ensuring equitable representation of all sections of people and their problems? (RQ-4 [Chapter 7]) How and by whom are policies justified through the mass media in India? 12

2.3 High level overview of the data in the social network graph database (knowledge base): the timed and untimed (static) relations are shown in edge labels. 18

2.4 Snapshots of the GEM website 28

4.1 Indicator plot for the four years with rank normalization: the blue curve denotes I_{CP}, and the box plots denote I_{rand} 64

4.2 CDFs of degree centralities of interlocking bureaucrats and politicians . . . 68

4.3 Change in clustering coefficient of the corporate-government network with time 68

5.1 [RQ2] Euclidean distance of relative coverage and mean relative coverage (across news-sources) for the four policy events. Higher the deviation for a particular news source, more different is its coverage from the mean behavior across news-sources. 84

5.2 PCA on constituency vectors for the four events: Principal component PC1 represents news-sources that cover more of informal sector, poor, and middle class (towards right) related issues, and political or corporate related issues (towards left). Principal component PC2 represents news-sources that cover political, corporate, and informal sector related issues (on the negative side). 87

5.3 Deviation of relative coverage of entity groups from their mean relative coverage across news-sources. Mean coverage is taken as the average coverage of an entity group across all news-sources. 88

5.4 CDF plot of article sentiment and tweet sentiment for the set TweetFol, for *The Hindu*. For the other news-sources for all events, we present the results in the Appendix. 91

6.1 Relative aspect coverage of each policy by mass media, social media community, and QH data 100

6.2 Relative coverage of aspects provided by political parties in QH data for the four policies 105

7.1 Plot of the relative coverage of top 20 entities for each policy for statements made by them: relative coverage is calculated as the number of statements made by the entity divided by the total number of statements by all entities, corresponding to a policy. 117

7.2 Plot of the aggregate sentiment, color coded on *degpol* for the top 20 entities with highest coverage: the aggregate sentiment/degpol is calculated as the sum total of the values corresponding to the statements made by an entity. Higher the value of degpol (darker the color of the bar), more is the overall polarity. 121

7.3 Mean relative coverage of aspects corresponding to the four policy events. . 125

8.1 Comparison of the relative aspect coverage of Google Alerts and that of news-sources, showing a strong similarity in the aspect coverage trend followed by the two (cosine similarities indicated in boxes within the plots). The bias in aspect coverage is also evident in both of the sources. Further, Pearson Coefficient for the four cases are 0.92, 0.6, 0.6, 0.76, respectively. . 134

8.2 Architecture of our news recommendation framework 136

8.3 Evolution of news-feeds over time: we consider an event with three aspects using which daily feeds are produced by our algorithm. The aspects are selected for exposure in descending order of the corresponding loss as indicated by the width of the aspect in a feed (more the width, greater is the number of articles displayed from that aspect). Each time an aspect is exposed in a feed, its loss diminishes as A_j gets closer to D_j. The algorithm stops exposing an aspect when the loss is reduced to zero. 141

8.4 Heat map for combination of fairness, diversity, and recency corresponding to the four policies: the area with red borders indicate the zones where the algorithm performs decently in terms of fairness, diversity, and recency. The optimal values of fairness and diversity coefficients are chosen as (0.5, 0.8). 148

8.5 GINI and HHI Plots for Aadhaar and GST: our algorithm is seen to outperform all of the baselines for its optimal combination of parameters $(f* = 0.5, d* = 0.8)$ 150

8.6 Weekly average news-feed age for our recommendation algorithm and the baselines 151

8.7 Repetition-at-k plots for the baselines and our algorithm, for all the four policies. *Note that for Google Alerts we do not plot for* $k = 0$ *as we do not have knowledge about the whole corpus of news articles from which it selects news. Thus, we do not know which articles it does not alert us about.* 153

A.1 Aggregate relative coverage provided by the mass media and its social media follower community corresponding to each policy: the blue bars and red bars represent mass media and social media coverage, respectively. . . . 200

A.2 Relative coverage provided by the mass media to each of the five constituencies for Demonetization and Farmers' Protests 204

A.3 CDF plots of article sentiment and tweet sentiment for the set TweetFol, for Demonetization, across news-sources. 213

A.4 CDF plots of article sentiment and tweet sentiment for the set TweetFol, for Aadhaar, across news-sources. 214

A.5 CDF plots of article sentiment and tweet sentiment for the set TweetFol, for GST, across news-sources. 215

A.6 CDF plots of article sentiment and tweet sentiment for the set TweetFol, for Farmers' Protest, across news-sources. 216

B.1 Comparison of retraining and inferencing schemes with respect to the gold model considering k=2 months: The positive percentage settles at around 70% for both of the schemes 219

B.2 Direct comparison of retraining and inferencing schemes: we consider a threshold positive percentage of 88% to define the fairness window, since this gives us an acceptable time period after which we can retrain the model. The minimum period for which a positive percentage>88% is maintained across Aadhaar, Demonetization, and GST turns out to be three months. 220

B.3 Relaxing the 15-day criteria for selecting articles in an attempt to pick under-represented aspects as suggested by U_j scores tends to significantly worsen the average feed age. 223

B.4 GINI and HHI plots for Aadhaar+Demonetization for pro and anti policy articles. For the optimal parameters of $(f* = 0.5, d* = 0.8)$ our algorithm outperforms the two baselines in terms of fairness and diversity. 226

B.5 Weekly average news-feed age for our recommendation algorithm and the baselines, for Demonetization and Farmers' Protests 226

B.6 GINI and HHI Plots for Demonetization and Farmers' Protest: our algorithm is seen to outperform all of the baselines for its optimal combination of parameters ($f* = 0.5, d* = 0.8$) 227

List of Tables

1.1 Applications and the corresponding research questions 7

2.1 List of manually collected keywords used to extract articles (and tweets) corresponding to the economic policy events. Here, we only show the manually selected keywords after converting them to lowercase, and after preprocessing of the articles was done. 15

2.2 Performance of ER within the media database for the person and non-person (Object) entities: there is an overall improvement due to context enhancement over time. 20

4.2 Count of bridge edges added during each time period (untimed edges were considered for calculations across all time periods). POL, COM, BoD, IAS stand for politicians, companies, directors, and bureaucrats respectively. The Govt/Public links are for appointments of bureaucrats in state owned companies, and are not considered in the calculations. 67

4.1 Overview of web data collected for the corporate-government knowledge base 71

5.1 Relative aspect coverage for mass media and social media, for the top five highest covered aspects in mass media 83

5.2 [RQ3] JS divergence showing difference in aspect coverage between mass media and social media: for TeleG, we could not find any tweet for Demonetization and GST. The Kolmogorov-Smirnov 2-sample test also suggest that the aspect coverage are significantly similar between the mass media and social media. 90

5.3 Odds-ratio of overlap of follower community for each pair of news-sources . 92

7.1 List of manually collected keywords used to extract articles (and tweets) corresponding to the ICTD policy events. Here, we only show the manually selected keywords after converting them to lowercase, and after preprocessing of the articles was done. 110

7.2 Relative coverage in percentage for entity groups (considering both *about* and *by* statements): BJP and INC are the two biggest parties in India (BJP being the ruling party currently). 120

A.1 KS statistics (2-sample test) for relative coverage provided by the mass media to the five constituencies for Demonetization. All p-values lie below 0.05. 205

A.2 KS statistics (2-sample test) for relative coverage provided by the mass media to the five constituencies for Aadhaar. All p-values lie below 0.05. . 205

A.3 KS statistics (2-sample test) for relative coverage provided by the mass media to the five constituencies for GST. All p-values lie below 0.05. 205

A.4 KS statistics (2-sample test) for relative coverage provided by the mass media to the five constituencies for Farmers' Protests. All p-values lie below 0.05. 206

A.5 Snapshot of the coding schema for Demonetization 208

A.6 Alignment matrix for Demonetization 209

A.7 Alignment matrix for Farmers' Protests 210

A.8 Alignment matrix for Aadhaar 211

A.9 Alignment matrix for GST 212

Chapter 1

Introduction

The purpose of this research is to build tools and techniques to monitor some aspects of the political economy around various policies, to provide a better and nuanced understanding of the policy process to researchers, journalists, and citizens. Political economy analysis studies the interaction between political and economic processes in a society. Specifically, it analyses the distribution of power and wealth between different entities or groups of entities, and the factors which create, sustain and transform these interactions. Overall, it helps provide a lens through which policies and the policy-making processes can be critically examined and questioned. The broad research question we ask in this thesis is: *Can web data be used to understand some aspects of the political economy around key policies in India?*

Policies are shaped by the political economy in a country [58], and significantly impacts its development. The tools developed through this research work can provide cleaner and complete information to the people about the political economy, thereby helping them to recognize its importance in development. This can also help to understand how different participants in a democracy (like the mass media, the citizens through social media, the business corporations, and the Parliament) may be aligned with one another.

There are several factors that can adversely impact development through incorrect or inappropriate policy formulation and implementation. These factors often arise due to the lack of proper functioning of the participants in the democratic process. For instance, a biased media leads to a skewed public opinion, resulting in elections not being contested on well informed policy grounds. Similarly, incorrect information propagating through information channels like mass media and social media can also lead to the citizens being misinformed on the current affairs. This may result in undeserving candidates securing key positions through the electoral process, leading to improper policy formulation and implementation. Corruption between the corporate and government entities can lead to corporate entities influencing policy, to skew it in their favor. Lack of nuanced understanding of the policy requirements can also lead to superficial parliamentary discourse on the citizens' concerns. My research touches upon each of these factors, by analyzing data collected from publicly available web based sources.

Political economy analysis is a wide field of inquiry that employs several methods of data analysis. We have worked on large-scale mass media data, and other publicly available data on corporations, politicians, bureaucrats, judiciary members, industry bodies, and civil society members to analyze a few areas that fall under political economy analysis, which include media bias analysis; measuring representation of different constituencies of citizens in the mass media, social media, and the Parliament; identification of people or organizations mentioned in the policy discourse who may be shaping policy; understanding the ideological positioning of these entities; and observing the evolution of interconnections between these entities over time. We use standard data analytical methods like topic modeling, sentiment analysis, factor analysis, and social network analysis alongside qualitative content analysis on this data to draw inferences on these areas, corresponding to some economic and technology policies in India. Next, we discuss the sub-questions corresponding to the broad research question that we answer through this thesis.

1.1 Research Questions

We divide the broad research question, *Can web data be used to understand some aspects of the political economy around key policies in India?*, into various sub-questions. The first question we attempt to answer is about corporate-government interlocks that influence the policy process: *(RQ-1) How can corporate-government interlocks be identified that may have a potential influence on the policy process?* In our work, we study the relationships between key government stakeholders and big business houses. We crawl web data from a large variety of sources, and extract relevant entities and relationships to construct a knowledge base of important policymakers and corporations in India. We also use network theoretical approaches to study the weights of the interlocks or overlaps between these stakeholders, and observe their evolution over time using an empirical indicator that we define later. We find that most of the interlocking nodes increase their connectivity (degree) over time, leading to the increase of interlocks between corporate and government entities over time. This indicates formation of a power structure among the influential politicians, bureaucrats, and business-persons, which may have a potential influence on policy-making.

The second research question that we answer is: *(RQ-2) Is the mass media biased in how it represents different policies?* To answer this question, we analyze mass media data to identify the coverage bias in representation of different aspects and constituencies pertaining to a policy. We also see the bias that individual media houses show in terms of coverage of the major political parties in India, which gives us an indication towards the biases that these houses carry with respect to their political favoritism.

The third research question that we study is: *(RQ-3) Is the policy-making process democratic, i.e., it ensures equitable representation of all sections of people and their problems?* To answer this question, we study the content discussed on policy matters in the three participants of democracy, namely the mass media, the social media, and the Parliament. As discussed earlier, we identify the aspects in the policy discourse using automated clustering techniques like Latent Dirichlet Allocation (LDA), and map these aspects to five

dominant frames or *constituencies* through which the content is presented. Finally, we analyze the biases in coverage of these aspects and constituencies, which tells us if the policy discourse is representative enough towards all sections of people and their issues of concern. We find that the concerns of all sections of people are not equitably covered by the three participants of democracy. Some of these sections of people (especially the poor) do not receive significant coverage in the policy discourse. In case they do, it happens mostly as a by-product of politicization of the issue at hand. We also observe that the structural issues of concern pertaining to the policies are generally neglected, and the policymakers rather focus on short-term, quick solutions of the problems.

The fourth research question that we try to answer is: *(RQ-4) How are policies justified through mass media to the citizens in India, and by whom?* We do this by employing different techniques that span the areas of natural language processing, information retrieval, machine learning, and data analysis. We use automated crawlers to crawl news websites on a daily basis. We then use an entity resolution heuristic developed by us, which uses edit distance, phonetics, and context based resolution approaches to resolve entities extracted from this large scale, unstructured data. We also use NLP techniques like dependency parsing and sentiment analysis to understand the stance taken by key stakeholders pertaining to these policies in mass media. To analyze the dominant topics of discussion (aspects) in the policy discourse, we use clustering techniques like Latent Dirichlet Allocation (LDA) and sentiment analysis. We find that policies are mostly justified by policymakers in mass media, who speak in favor of the policy if it has been formulated and implemented during their tenure, and oppose it otherwise. Next to politicians, business-persons receive the maximum coverage in mass media, and similar to the politicians, they speak positively about the policies, primarily focusing on the technical advancement that they bring. In comparison, policy experts, authors, social activists, and academicians receive insignificant coverage on mass media in the policy discourse.

Finally, we also describe our initial attempt in addressing the issue of algorithmic bias in recommendation algorithms suggesting news feeds. Here we ask the research question: *(RQ-5) Can we produce a news-feed that is unbiased and fair, in terms of its representation*

of news? We study the coverage provided to different aspects by Google Alerts, a popular content change detection and notification service, and suggest our own algorithm for designing a *fair* and *balanced* news aggregator, which provides a fair representation to different aspects under discussion in the policy discourse.

1.2 System Architecture

Using the aforementioned techniques, we have developed a technological platform that enables us to study any policy, with respect to the research question described in the previous section. The block diagram in figure 1.1 provides a brief overview of our system architecture. Our system consists of three operational layers namely, the *collection* layer where data is crawled from a wide variety of web based media sources and sources of corporate-political data; the *Algorithms* layer where we apply different algorithms to analyze the data; and the *Applications* layer where we build applications to present the findings of this analysis. One of these applications, the *Topic Analyzer* studies the topics or aspects being discussed in the policy discourse, and also analyses the bias present in the Indian mass media, social media, and parliamentary question hour data with respect to policy representation. The *indicator monitor* application tracks changes in the corporate-government interlock patterns to find instances noteworthy of further investigation. Finally, the *News Feed* attempts to provide a fair and balanced news feed to the user, which counters media bias algorithmically. Table 1.1 shows the applications and the corresponding research questions that they help us answer.

This work is structured as follows: In the next section, we position our work with respect to the other papers in political economy analysis. We then describe our research method in chapter 2. In chapter 3, we describe some of the relevant studies related to our analysis. Chapter 5 describes the study of political economy around policies in terms of study of bias in Indian mass media. Chapter 6 studies if policy-making in India is democratic, and provides equitable representation to the concerns of all sections of people. Chapter 7

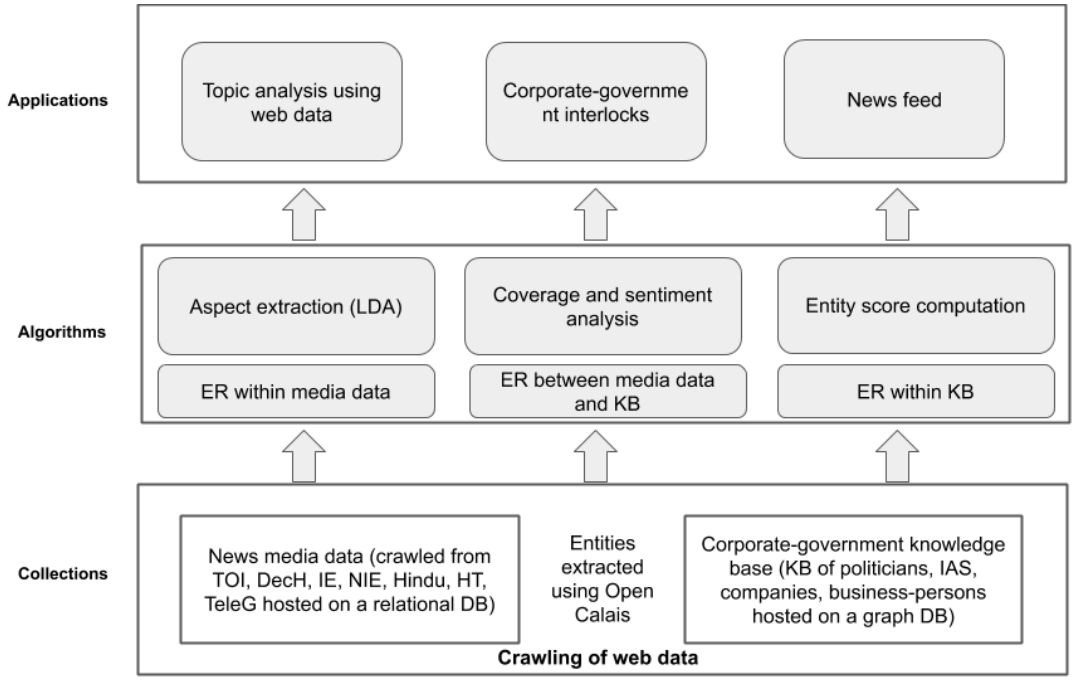


Figure 1.1: Overall architecture of the system: The lowermost layer belongs to data collection where web data is crawled from a multitude of sources; the middle layer contains the algorithms used to clean and analyze the data; the uppermost layer contains the applications that we build upon this analysis.

Application Name	Research Questions
Indicator Monitor	(RQ-1) How can corporate-government interlocks be identified that may have a potential influence on the policy process (chapter 4)?
Topic Analyzer (of Web Data)	(RQ-2) Is mass media biased in how it represents different policies (chapter 5)? (RQ-3) Is the policy-making process democratic, i.e., it ensures equitable representation of all sections of people and their problems (chapter 6)? (RQ-4) How are policies justified through the mass media to the citizens in India, and by whom (chapter 7)?
News Feed	(RQ-5) Can we produce a news-feed that is unbiased and fair, in terms of its representation of news (chapter 8)?

Table 1.1: Applications and the corresponding research questions

discusses how policies are represented in Indian mass media, and the ways in which this representation is biased. Chapter 4 talks about the indicator of corporate-government overlap that we have developed, and how corporations have a potential influence in policy-making. Finally, in chapter 8 we discuss our initial plans of developing a fair and balanced news recommendation system.

1.3 Thesis Structure

In this section, we briefly describe how this thesis is organized. In the chapter 2, we describe the Research Methodology followed for the analysis of political economy around some key policies. This includes the technological system that we have built for political economy analysis, followed by the qualitative content analysis that we perform on the data collected. Finally, we discuss the website that we have built to present our findings.

Chapter 3 discusses the relevant works done in this domain. It includes discussion on studies related to political economy analysis, followed by the roles played by the mass

media and social media in impacting political behavior of citizens, followed by studies on analysis of bias in these sources of information. We also look at a few studies on policy representation in the Parliament, which is another major stakeholder of any democracy. Finally, we conclude with the caveats and loopholes that one needs to look out for while doing big data analysis, which forms a chief component of this thesis.

Chapter 4 deals with the study of corporate-government interlocks. Here, we discuss how we rank the elites (entities of importance) based on their centrality, followed by the analysis of interlocks or connections between them. We also discuss *indicator monitor application*, which is an application that quantifies the extent of interlock between corporate and government entities in India, and also studies its evolution over time. We conclude with a discussion on the causes behind these interlocks.

In chapter 7, we study the representation of some key policies in the Indian mass media. We study seven news-sources of importance, and find out which entities speak the most on these policies, how they speak on them, and which aspects from these policies are preferentially covered by these news-sources. This gives us some idea about the political economy around the policies considered.

While political discourse depends a lot on what the policymakers and other influential entities talk about in the media, it is also a fact that mass media itself might be biased in its content presentation. Some aspects of a policy event might be over-represented by the media compared to the others. In chapter 5, we study the biases created in the Indian mass media through the effects of *agenda-setting* and *framing*. Specifically, we study the bias of the media towards or against some aspects, and the frames through which these aspects are presented to the audience. We also see if the social media in any way counters these biases and acts as an independent information source, or simply echoes what is being discussed in the mass media, further amplifying the biases.

In chapter 6, we dig deeper into the aspects being discussed in the three prime participants of democracy – the mass media, the social media, and the Parliament of India – in terms of the four economic policies we consider. We analyze the statements made on the various

aspects of these policy issues, and attempt to find out if they differ in any way from each other, with respect to the discourse on issues relevant to the different sections of people.

While in the aforementioned chapters we study how the policies are represented in the participants of democracy, and if there exist biases in their content, we also want to see if there exist inherent biases in the algorithms that recommend news to the users. In this direction, we propose an algorithm that achieves fairness and diversity in content representation relevant to policy issues, and discuss our initial experiments towards developing a fair news aggregator (chapter 8).

Finally, in chapter 9, we discuss some of the technical challenges that we faced during the course of this work, and the limitations of this study. We conclude with the primary findings of our research, and describe some of our future goals in this direction.

Chapter 2

Research Methodology

The diagrams in figure 2.1 and 2.2 describe our research methodology pipeline. We have built a technology platform that produces processed data obtained from different data sources – data collected from mass media websites and other web based sources. The processed data includes but is not limited to aspects present in the policy discourse in mass media, resolved entities obtained after entity resolution, sentiment slant of articles and statements made by these entities about policies, and the corporate-government interlock data. We also carry out qualitative content analysis of this processed data that includes studying news articles belonging to various aspects to name the aspects, studying dominant frames in which news is presented, and understanding in detail how these aspects and frames are represented in popular media. Finally, the results from our qualitative analysis and the technological system are collected presented in a website, which is updated periodically. The website is intended to aid us in disseminating our research and analysis to our target audience, which includes journalists, social scientists, social activists, and researchers. While the pipeline presented in figure 2.1 helps us answer the research question related to corporate-government interlock, the pipeline in figure 2.2 addresses the rest of the research questions.

Our research method follows this pipeline to answer the research questions explained

in the aforementioned sections. Analysis of the research questions on mass media bias, justification of policies by entities in mass media, and representativeness of policy-making in the three participants of democracy require the three stage pipeline. For the research questions on corporate-government interlocks, and fairness in news presentation, we do not perform much of qualitative analysis. Hence, to answer them our research method follows a two stage pipeline of data processing and data presentation. In the subsequent sub-sections, we describe our research method in further detail.

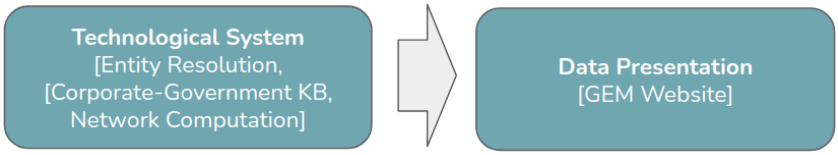


Figure 2.1: Pipeline to answer: (RQ-1 [Chapter 4]) How can corporate-government interlocks be identified that may have a potential influence on the policy process?

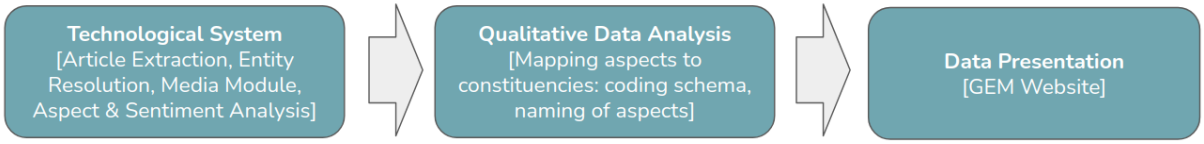


Figure 2.2: Pipeline to answer: (RQ-2 [Chapter 5]) Is mass media biased in how it represents policies? (RQ-3 [Chapter 6]) Is the policy-making process democratic, i.e., one ensuring equitable representation of all sections of people and their problems? (RQ-4 [Chapter 7]) How and by whom are policies justified through the mass media in India?

2.1 Technological System

Data is collected by our technological system from a wide variety of web sources including websites of popular news-sources, websites of corporate and government data, and other

knowledge repositories. The system consists of two infrastructural components, using which it collects, stores, and analyses this data – the corporate-government knowledge base and the media module. We describe the technological system in detail in this section.

2.1.1 Data Collection

The technological system crawls data from different pre-identified structured and unstructured web sources longitudinally, to create and update *corporate-government knowledge base*. This data now includes information about politicians who contested the national elections from 2004, all chief ministers of states over time, retired and current bureaucrats (officers of the Indian Administrative Services), a snow-balled network of companies and their board of directors starting with approximately 5000 public listed companies in India, their subsidiaries, business-persons (board of directors (BoDs) and executive management of the firms), government departments, ministries, and family members of the business persons and politicians. The snow-balling of the company network was done up to a depth of three to include subsidiaries of companies and companies connected through co-occurring members in their board of directors. This network therefore contains the largest companies and most important politicians and business-persons, in line with our goal to be able to examine corporate-government interlocks among the big corporate houses and prominent politicians. We host the knowledge base in Neo4j [209], which is a graph database.

Each of these entities are stored in the form of nodes in the knowledge base (KB), which also contains relations (edges) between these entities in the form of edges. The relations in our dataset are of two types: *explicit* and *implicit*. Explicit relations are relations that are directly observable and for which a direct evidence exists (for example, the connection between a politician and a company through board membership). On the other hand, implicit relations often include connections that do not have any direct evidence, but can be assumed implicitly. For example, a company and a politician may be assumed to have an implicit connection, if the company has contributed to CSR activities in the

politician's constituency, and the politician is a minister or an MP. Here, although we do not have any direct evidence of the connection, we can label such a link as an implicit interlock to indicate the possibility of a real relationship. Both nodes and edges have multiple properties associated with them in the KB.

Before storing the data, we resolve the entities present in the raw data using the properties associated with them. The entity resolution algorithm (ER) is described in section 2.1.2. We run network computations (a modified version of PageRank) on this KB to obtain a ranked list of entities of each type, based on different patterns of corporate-government interlocks. These ranks are then used to compute an *indicator of interlock*, using which we study the weight of the interlocks between the corporate and government entities, and the evolution of these interlocks over time.

The system also crawls news articles on a daily basis from seven popular newspapers (*The Hindu, The Times of India, Indian Express, The New Indian Express, Telegraph, Deccan Herald* and *Hindustan Times*) in India from 2011. The news data (news articles with other meta-data) is stored in the *media module*. News articles belonging to categories like national, international, regional, sports, opinion and business are collected, along with the URLs and their meta-data and stored in an *article database* within the media module. Articles belonging to various policy events are then filtered out from this database using an augmented keyword based approach. Articles are initially collected based on a set of manually selected seed set of keywords related to the event[1] as shown in table 2.1. Once the initial set of articles are retrieved from the media database, the seed set of keywords is augmented with newer keywords extracted from these articles based on their relevance scores, to be used for extraction of other articles based on this augmented keyword set. In each iteration, more keywords are filtered out, and a score of relevance is obtained corresponding to each extracted keyword. Finally after saturation, only the top 20% of the keywords with highest relevance scores are considered as the final keyword set based

[1]We study seven policy events in total, which include economic and technology (ICTD) policies. Here, we show the keywords corresponding to the economic policies, since a majority of our analyses are based on them. The keywords for the ICTD policies are shown in chapter 7.

on which the articles are collected[2].

Keywords (manually selected)
Demonetization: demonitisation, demonitization, denomination note, cash withdrawal, swipe machine, unaccounted money, withdrawal limit, pos machine, fake currency, digital payment, digital transaction, cash transaction, cashless economy, black money, cash crunch, currency switch, long queue, demonetised note, cashless transaction, note ban, currency switch
Aadhaar: aadhar, aadhaar, adhar, adharcard, aadharcard, aadhaarcard, uidai, aadhar card, public distribution system, pds, ration card, ration, e-pos
GST: gst, goods and service tax, goods & services tax, gabbar singh tax, goods service tax, goods and services tax
Farmers' Protest: farm loan, crop loan, farmer suicide, debt waiver, waiver scheme, farming community, farmer agitation, plight farmer, distressed farmer, farmer issue, farmers protest, farmers' protest, agrarian crisis, agrarian unrest, farmers protests, farmers' protests, loan waivers, loan waiver, agriculture protest, farmers' march

Table 2.1: List of manually collected keywords used to extract articles (and tweets) corresponding to the economic policy events. Here, we only show the manually selected keywords after converting them to lowercase, and after pre-processing of the articles was done.

The system also runs a named entity recognition tool [167] to store the named entities with their properties – entities of type Person, Company, Organisation, City, and Province are extracted. Next, ER is performed on these entities where the entities are resolved against the previously resolved entities identified in older articles. Finally, we also match the resolved media entities in the media module with the resolved entities in the corporate-government KB. The number of entities present in the media data is of the order of millions, which makes ER within the media data a challenging task as each unresolved entity. We use an indexing solution [208] to make this efficient. The system then runs clustering (LDA) on this final article set for each event, to identify the aspects of discussion. It also uses dependency parsing to identify the entities speaking on the

[2]Articles where three or more of these keywords occur (the count also includes keywords that are repeated) are collected. In case, an article contains less than three keywords corresponding to an event, we see if the keywords occurring are actually talking about an event. If so, we store the article in the media module.

policies in these articles, and passes their statements through a sentiment analyzer to identify their stance with respect to the policy.

Additionally, we also have the social media (Twitter) data for the followers of the news-sources considered collected by the system. Tweets and retweets of all followers of a news-source handle are collected first. Next, the tweets corresponding to each policy event are extracted using the same set of keywords as shown in table 2.1. The number of followers for the news-sources are: TOI (11026374), HT (6299716), Hindu (4842234), IE (2742132), NIE (347148), TeleG (51884), and DecH (24896). The number of tweets for the economic policy events are 396499 for Aadhaar, 1236500 for Demonetization, 1147154 for GST, and 512457 for Farmers' Protest. We refer to this set of tweets by the follower community of news-sources as *TweetFol* throughout the thesis.

We also take into consideration 135,460 questions raised during the period of 15th (2009 - 2014) and 16th (2014 - till date) LokSabha (lower house of the Indian Parliament) sessions. All elected representatives have the right of raising questions in the Indian legislature – Members of Parliament (MPs) ask questions to ministers eliciting information regarding different policies. The first hour of each day's parliamentary session is reserved for what is called the 'Question Hour'. The complete data for these questions is available on the LokSabha's website. We finally obtain 351 questions for Demonetization, 89 questions for Aadhaar, 151 questions for GST, and 140 questions for Farmers' Protest in the entire QH data for 15th and 16th LokSabha. The extraction of these policy specific questions is done using the the same set of keywords as mentioned in table 2.1. We manually map each of these questions to an aspect already created for mass media data. This mapping is done by two annotators after coming to a mutual agreement on each mapping, enabling us to compare the QH and mass media data, on the same aspects.

Note that the ruling parties were different in the two LokSabha sessions that we consider – while in the 15th LokSabha, the United Progressive Alliance was the ruling alliance (with Indian National Congress (INC) as the largest coalition party), in the 16th LokSabha, the Bharatiya Janta Party was the ruling party with INC in opposition. All of our findings on the parliamentary data hold for both of the parties and sessions.

2.1.2 Entity Resolution

The data for our analysis has been collected from multiple sources. The same entity might be mentioned in different forms in this wide variety of data, in terms of the spelling of its name, its abbreviation, and so on. For this purpose, we need an Entity Resolution (ER) heuristic to standardize the entity names. We perform ER at three stages – within the mass media database, within the corporate-government KB of entities, and between the mass media database and the KB of entities.

Entity resolution in the corporate-government KB: The ER process in the KB works in two steps: (i) matching using the context data for entities, and (ii) further filtration using string and phonetics based similarity measures on the entity names. Different pieces of context information are used in different cases based on the available data. To resolve newly crawled *politicians* with existing politicians, we use their political parties and political titles (like prime minister, chief minister, etc.) as the context information. To resolve them with business-persons and bureaucrats, we use the context information available in Wikipedia's *Category* section of the entities' pages[3]. *Bureaucrats* are resolved against business-persons using their date of birth and names. Finally, *Business-persons* are resolved among themselves using the *company identification numbers* (CINs) of the companies with which they are associated. To resolve *companies* with each other, data from some sources was straightforward to match based on the CINs, and in other cases context information was used such as the registered location of the company.

After context based matching, sometimes in cases where a precise identification can not be done, the candidate list of matching entities are resolved using string+phonetics based approaches on an experimentally decided threshold of similarity. We measured the performance of our ER approach based on randomly selected 2000 person entities, with equal number of positive and negative examples[4] annotated by two of the annotators. We ob-

[3]For example, if a politician has also been a bureaucrat at some point before her political career, usually *Indian Administrative Service* is mentioned in the category section of her Wikipedia page.

[4]Positive: new entities that can be merged with existing entities, Negative: entities that do not match and hence cannot be merged with existing entities.

tained precision and recall[5] values of 95% and 97.4% respectively (apart from companies (which could be mostly resolved by CINs), there are very few non-person entities. Hence, the ER performance for them were nearly 100%.), and managed to build a social network graph (knowledge base) with relations of the types: *politician—belongsto—political party (20951), subsidiary—belongsto—company (453)*, etc. Figure 2.3 provides a high level view of the social network graph data where these relationships are shown. Although our currently obtained relationships are all extracted from the network structured knowledge base, we are currently augmenting this with relationships extracted from mass media data, using NLP and *Snowball* based techniques [12, 234].

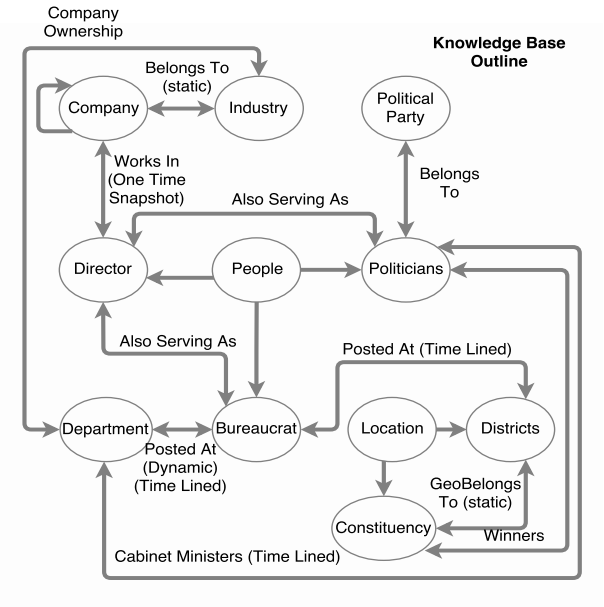


Figure 2.3: High level overview of the data in the social network graph database (knowledge base): the timed and untimed (static) relations are shown in edge labels.

Entity resolution within mass media: Since news articles are crawled continuously, we maintain a set of *resolved entities* which have been successfully resolved so far, and

[5]Throughout this work, we consider correctly resolved entities as TP, correctly unresolved entities as TN, incorrectly resolved entities as FP, and incorrectly unresolved entities as FN.

keep augmenting it as more news articles are crawled and throw up additional entities to be resolved. On encountering a new unresolved entity, ER within media data follows two steps: (a) find the top ten candidate entities from the resolved set of entities based on partial matching (match of at least one word) of their standard names, aliases, and context; and then (b) further filter these top ten entities to obtain a set of best matches, using string+phonetics based distance measures applied on standard names and context of entities, based on experimentally set similarity thresholds The ER process uses contextual attributes returned by the named entity recognition tool [167]. These include attributes like the type of entity, its standard name, and some other context information (especially for the non-person entities like latitude and longitude for locations). We also store the entities frequently co-occurring in the same articles as the entity to be resolved, based on their TF-IDF scores. We call such entities *associated entities*. We follow a strategy of merging this context information together for entities which are successfully matched with each other. This improves the accuracy of the resolver over time as it gains more and more context information for each resolved entity. If any of these steps fail, we make a new entry in the resolved entity set.

To evaluate the performance of ER within the media data, we collected 100 random entities from the last 50 news articles for every year from 2011 to 2017. For each time period, the 100 unresolved entities from these 50 articles are resolved against all of the remaining resolved entities whose context information has been enriched with data from the previous years combined. For example, entities from the last 50 articles within the 2012-13 period are resolved with the remaining resolved entities during 2011-12 and 2012-13. Table 2.2 shows the precision and recall values for each of these sets of last 100 entities. Although the values do not exhibit a strict monotonic improvement with increasing context, we can see that there is a significant overall improvement from the first time period (2011-12) to the last (2016-17) in both of the types of entities (Person and Non-person).

Entity resolution between the media database and knowledge base of entities: The media entities thus resolved, are next also resolved against the entities in the social network (Neo4j graph database). This process of ER between the media and the knowl-

Interval	Person		Non-person	
	Precision%	Recall%	Precision%	Recall%
2011-12	87.5	93.9	88.6	94.59
2012-13	83.52	95.94	81.17	97.18
2013-14	89.41	98.7	91.02	98.61
2014-15	91.95	95.23	86.58	100
2014-16	95.18	92.94	89.02	97.33
2016-17	97.61	96.47	93.82	96.2

Table 2.2: Performance of ER within the media database for the person and non-person (Object) entities: there is an overall improvement due to context enhancement over time.

edge base of entities starts with alias matching, and further filters the result set through context matching (an approach similar to the ER approach for the knowledge base). We also use *associated entities* for context matching, which are entities co-occurring in any article in the media data, or neighboring entities in the knowledge base data, and it is able to improve the ER performance significantly. The performance of ER between the knowledge base and media entities was also satisfactory. We randomly selected 50 person and 50 non-person entities and observed the precision and recall values as 93.18% and 95.35% respectively for persons, and 87.5% and 89.74% respectively for non-persons.

2.1.3 Aspect extraction using LDA

Discussion on a policy event often includes several aspects, issues, or topics of discussion. For example, *Long queues at ATM counters leading to plight of the middle class and poor* is one of the aspects for Demonetization that discusses the kind of problems people faced during the cash-shortage in ATM counters post implementation of the policy. In this section, we explain the automated aspect extraction process from the discourse on policies, which is one of the contributions of our work.

Mass media: Latent Dirichlet Allocation (LDA) is used to identify different aspects

within each event, similar to [232]. LDA is a statistical model that maps a set of documents to unobserved topics, which aids in clustering similar documents into topic clusters that can be manually examined and labeled. In our case, the documents refer to the media articles, which are mapped to different topic clusters, which we refer to as *aspects* henceforth.

To measure the accuracy of LDA aspect mapping, two authors randomly selected 20 articles from each aspect identified by LDA for each event (around 300 articles from each policy event in total), named the aspect based on these articles, and checked belongingness of each article to the aspect by reading the article text and coming to an agreement. The articles that did not belong to the aspect were considered to be the false positives. The authors then checked the total number of true positives (and false positives) among all the articles studied in the event, and calculated the aspect mapping accuracies. The accuracies of mapping for the economic policies are 85% for Demonetization, 96% for Aadhaar, 81% for GST.

Our method of evaluating the accuracy of LDA mapping handles any case of aspect distribution, i.e, whether there are very few aspects in an event or a large number of aspects, or whether a majority of the articles in an event belong to just a small set of aspects. This is because we start with the LDA identified aspects, and evaluate equal number of articles from each of these aspects. Thus, any skew in aspect distribution does not affect our accuracy metric, since the content is balanced across categories.

Social media: We map only those tweets that contain URLs of mass media articles to the aspects to which these articles belong, since tweets are concise and sometimes even grammatically incorrect, which makes it difficult to map them to specific aspects. Hence, the mass media URL carried by the tweet aids in mapping the tweet to one or more of the existing aspects as defined for mass media. Note that for each follower of a news-source, we also obtain tweets written by the user that may refer to one of other six news-sources as well, and thus we are able to observe actions related to any of the news-sources by the overall social media follower community of the news-sources. The number of tweets containing article URLs of the four economic events are 34521 for Aadhaar, 59489 for

Demonetization, 38073 for GST, and 22820 for Farmers' Protest, which constitute 8.7%, 4.8%, 3.3%, and 4.4% of the total number of tweets collected, respectively.

2.1.4 Sentiment Analysis

For sentiment analysis of the statements made by the entities, the system first uses dependency parsing [124] to classify the statements where the entity is mentioned in the media articles into two classes, namely the *by class* (containing statements made by the entities covered by media) and the *about class* (statements made by the media about the entities). The parser helps identify relations like *nsubj, nmod, amod, and dobj* which are used as features: whenever there is a statement by an entity, the entity occurs in an *nsubj* relation in the dependency graph; and when an entity is being spoken about, it occurs in any one of *nmod, amod, and dobj* relations.

We experimented with the different sentiment analysis tools provided in the iFeel framework [168], and finally settled with Sentistrength for sentiment analysis of articles, and Vader [84] for sentiment analysis of tweets and statements by entities (since Vader specifically was designed for analysis of sentiment for short text) based on their performances in terms of accuracy. Sentistrength [215] reports TPOS (positivity) and TNEG (negativity) scores for each article. TPOS score is in the range of 1 (not positive) to 5 (extremely positive), and TNEG score is in the range of -1 (not negative) to -5 (extremely negative). The aggregate sentiment for an article is calculated as the sum of TPOS and TNEG. To measure the accuracy of sentiment given by Sentistrength, for the 300 articles selected in the previous section to measure the performance of LDA mapping, the authors also assigned a sentiment score manually (ground truth). Articles' sentiment alignment was found to be 84% for Demonetization, 76% for Aadhaar, 75% for GST and 84% for Farmers' Protest. The evaluation of accuracy that was done on content, was nearly balanced across the categories positive and negative, and hence, the accuracy calculation does not suffer from data bias.

For each of the *by* and *about* sentences for each entity, the aggregate sentiment slant is obtained using the *Vader* sentiment analysis tool. Vader provides an intensity score in the range of -4 to +4 for each word in a sentence, and an overall normalized compound score in the [-1,1] range for the entire sentence. We add up the positive intensity scores of each word to obtain the overall positivity score (TPOS), and the negative intensity scores to obtain the negativity score (TNEG) for a statement or tweet. This tool takes into account exclamation marks, punctuation, Degree modifiers (such as intensifiers, booster words, or degree adverbs), bigrams, trigrams and emoticons, which makes it more suitable to use for sentiment analysis of article sentences. We also obtain the *by* and *about* aggregate sentiments for entity groups (for example, political parties, corporates, and academicians), by summing over the sentiment slant of statements for all entities in that group.

The sentiment classification accuracies were in the range of 75-79% for all policies for 250 statements manually analyzed by four authors for each policy. While calculating the aggregate sentiment score for an entity (sum of sentiment scores for all statements for an entity), the positive and negative values might cancel out, leading to a nearly neutral (close to zero) aggregate score. To capture the polarity in expression, we define another measure of sentiment polarity named the *degree of polarization (degpol)*, which is computed as:

$$degpol(s) = \begin{cases} \frac{TPOS+1}{|TNEG|+1} & \text{if TPOS} \geq |TNEG| \\ \\ \frac{|TNEG|+1}{TPOS+1} & \text{if TPOS} \leq |TNEG| \end{cases}$$

degpol provides a measure of how polar the statement s is in terms of the sentiment polarity of the sentences in which the entity is mentioned. A non-polar statement will have the minimum *degpol* score of 1. We calculate the *degpol* of an entity as the sum of *degpol* of all statements made by the entity.

2.1.5 Computation of Entity Scores

Our knowledge base of entities has different types of links between entities as shown in figure 2.3. These links and the paths formed by these links are termed *interlocks*. To distinguish important interlocks from not-so-important interlocks, we wanted to rank the entities in the knowledge base graph on various patterns of interlock. The computations are performed in an offline manner and can be triggered manually upon any significant data update in the knowledge base graph.

We rank the entities based on their involvement in different types of patterns: *corporate connected politicians*, *corporate connected bureaucrats*, *politically connected firms*, and *politically connected managers*. The rankings are done through a series of PageRank computations for each pattern to obtain rank scores as a measure of an entity's involvement in instances of a pattern. Thus, the pattern *corporate connected politicians* is used to rank politicians based on their corporate connectedness, and so on. The details of this process are described in chapter 4.

2.2 Qualitative Analysis of Data

The technological system produces processed data required for our analysis. To obtain useful insights from this processed data, we need to perform some qualitative analysis on this data. In this section, we describe this analysis in detail.

Aspect naming: As discussed earlier, we use LDA to identify aspects corresponding to the policy events. While LDA provides us a set of coherent aspects, in order to understand what kind of articles these aspects contain, it is important to name these aspects. Aspect naming is also required since we need to map these aspects to the five dominant constituencies or frames of presentation. Three annotators studied 50 articles manually from each aspect per policy event to name the aspect, after coming to an agreement. The aspect names help us understand the issue that most articles belonging

to the aspect are talking about. To reduce subjectivity of the exercise, we finalize an aspect name only after all three annotators agree on it.

Constituency identification and mapping: One of the goals of our work is to identify the alignment of a news-source in terms of some standard constituencies, to study the framing effect. We identify five constituencies: *poor*, to provide for the poor typically through wealth distribution strategies; *middle class*, typically the middle class consumers who have disposable income, to benefit through tax breaks, lower prices, and use of technology; *corporate*, driven by big corporates and formalization, economic growth, free-market policies, minimum governance; *informal sector*, driven by small enterprises and aided by slow formalization of industries and trade including agriculture; and *government* in terms of pro/anti viewpoints towards the state.

For each policy event, we map each aspect to these five constituencies based on whether the aspect supports or opposes or is not applicable to the particular constituency. For example, for the *Demonetization* policy, the aspect on *Queues at banks and ATMs* is classified as pro-middle class because most articles on this aspect were negatively writing about the problems caused to the common people in getting cash at ATMs. The same aspect is classified as anti-government because negative articles on these aspects generally criticize the government's apathy and lack of foresightedness in handling the issue. These five constituencies and the news-sources' alignment to them help us in studying the effect of framing in mass media.

The aspect to constituency mapping was performed by three annotators, each of whom went through around 3000 articles in total (50 articles from each aspect for each event). A coding scheme was developed through this qualitative analysis to do this mapping (snapshot provided in the Appendix). We evaluated the inter-coder agreement for the coding exercise using the percentage agreement calculation method as described in [200]. For each policy event, we consider each (aspect,constituency) combination as a sample point, which is rated -1/0/+1 by three annotators. Based on the coding scheme, the initial mapping exercise had an inter-coder agreement of 61.33% for Demonetization,76% for Aadhaar, 71% for GST, 74.3% for Farmers' Protests. We ran another round of moderation

and due deliberation before finally coming up with the agreed upon coding scheme. Our method of building the coding scheme after rigorously going through the news articles by multiple annotators ensures that there is minimum bias in the mapping of aspects to constituencies. Further details are present in chapter 5 and in the Appendix.

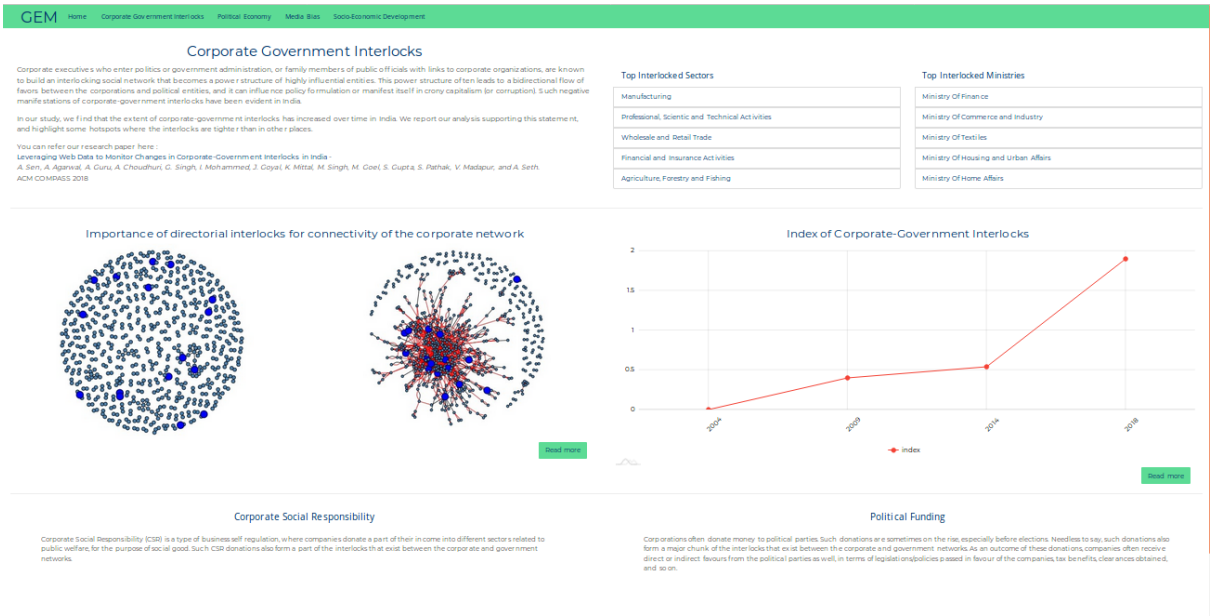
It is important to note that other kinds of constituencies can be added by identifying different frames. We chose the aforementioned frames because the specific economic policies considered in the paper are strongly related to these frames of content presentation and perception. It is also important to mention that this process of mapping aspects to constituencies can be extended to any dataset – although the size of the entire data analyzed in this work of the order of millions, the aspect to constituency mapping has been done manually by studying just 3000 articles, i.e., randomly selected 50 articles from each aspect for each event (there are around 15 aspects for each event).

Analysis of the coded content: To answer the various research questions discussed, we need to analyze the content coded in the aforementioned steps. We analyze the coded content in three ways: (a) We analyze the relative coverage provided to different aspects by the news-sources. Since the aspects are mapped to the dominant constituencies or frames in which news is presented, this also aids us in understanding which section of the society is over or under-represented in the news-source. (b) We analyze the sentiment slant corresponding to the policy discourse for each of these aspects. This helps us understand how the news-sources vary from each other in presenting these aspects to the public. Sentiment analysis also helps us in understanding the statement bias that the news-sources carry. Finally, using the coverage of aspects and their sentiment slant, we map the news-sources to a 5-dimensional constituency space, which helps us understand the extent to which a particular news-source varies from the average trend of presenting the constituencies. While conformist news-sources will present the constituencies similar to the average trend with which they are presented, the non-conformist news-sources will show variation from the trend. This analysis helps us understand media bias, where we see the different ways in which a news-source is biased towards or against certain aspects and constituencies.

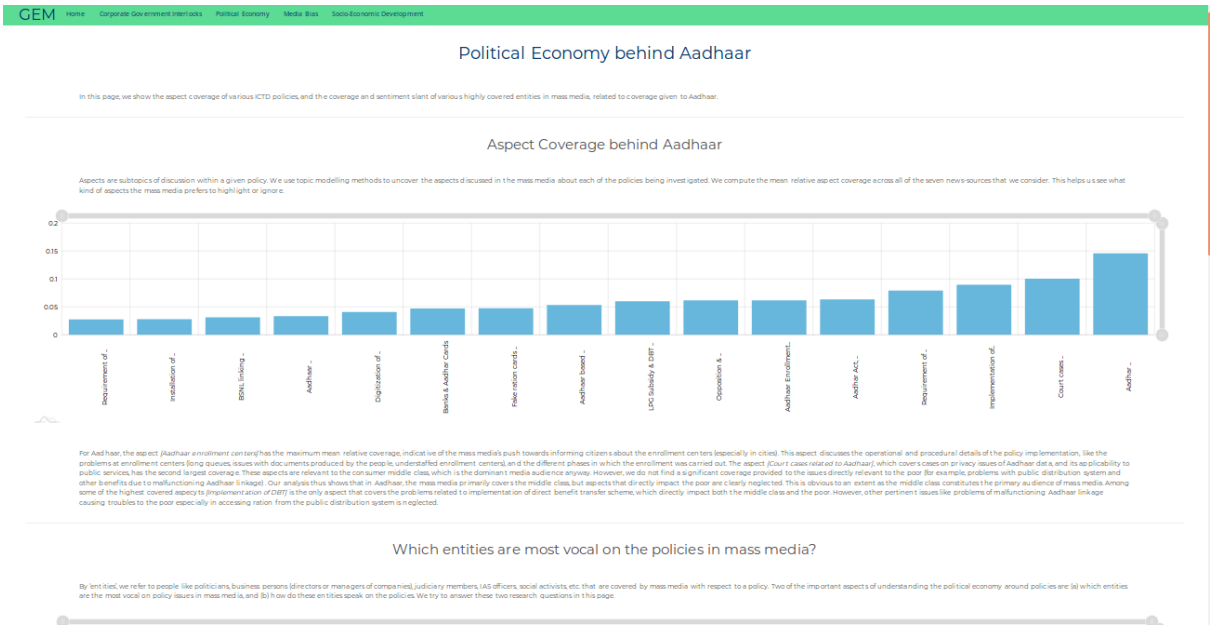
Our prime contribution lies in putting the entire process of automated and qualitative analysis into a framework, which includes a major part of technological intervention, and some amount of manual analysis. While the manual qualitative content analysis is important, it does not impact the generalizability of our method, which can be extended to other policies and even other geographies.

2.3 Data Presentation

We have developed a website to present the results of our data analysis. We call this site the *Giant Economy Monitor* (GEM)). It currently contains information and visualizations about all of our findings. The purpose of this website is to make our research reach the target users of our system. The primary target users of our system are the journalists, social activists, policy experts, and policymakers, who have an impact on the policy-making process. The website is supposed to serve as a one-spot repository of information about the policy-making process in India, which can eventually aid the target users understand and obtain feedback on the process of policy-making and on how these policies are represented to the citizens. Currently, our website has four major components or tabs namely, the *Corporate-Government Interlocks* tab, the *Political Economy* tab, the *Media Bias* tab, and the *Socioeconomic Development* tab. These tabs elaborate the results obtained from the application layer of our system. Figure 2.4 shows some snapshots of our website. Eventually we want to find better ways in which we can present this information to people, to have them critically reflect on the political economy.



(a) Corporate-Government Interlock Tab



(b) Political Economy Tab (entity coverage)

Figure 2.4: Snapshots of the GEM website

Chapter 3

Related Work

In this chapter, we touch upon the related works that have motivated this thesis. We start with a few foundational studies on political economy analysis, followed by some studies on applications, frameworks, and methods of political economy analysis. We then focus on how the three active participants of democracy – the mass media, the social media, and the Parliament – impact policy-making, followed by some works that study the biases existing in these participants. We finally end this chapter with discussion of some commonly discussed caveats of big data analysis.

3.1 Political Economy Analysis and Its Importance

3.1.1 Foundational Studies on Political Economy Analysis

Political economy analysis was developed as a field of intellectual enquiry in the 16^{th} to 18^{th} century of the mercantilist school that called for a strong role for the state in economic regulation. Scottish economist Sir James Steuart, 4th Baronet Denham, whose work regarding the Principles of Political Economy [201] is considered the first systematic

work in the area. Denham stated that the government has a key role in the economic development of society, particularly in the management of population and employment. According to him, government intervention was also required to bring market equilibrium. Subsequently, studies by philosophers like Adam Smith [192] attributed wealth distribution to political, economic, technological, natural, and social factors and the complex interactions between them. In the 19th century, political economist David Ricardo further developed Smith's ideas [170], and posited that states should produce and export only those goods that they can generate at a lower cost than other nations, and import goods which other countries can produce more efficiently. His work extolled the benefits of free trade. In the mid-19th century, Karl Marx proposed a class-based analysis of political economy [126], and analyzed the functioning of the capitalist system. Marx stated that capital's starting point is circulation of commodities, whose ultimate product is money. According to Marx, the capitalists operate solely to increase their wealth, and they indulge in purchase or facilitating production of commodities primarily to generate surplus value by selling these commodities. In the process, they end up exploiting the working class by depriving them of their deserved wages (value for labor).

The more recent studies focus on how the capitalism based neo-liberal economic system aids in accumulation of a disproportionate amount of wealth for the select few 'economic elites'. David Harvey, in his seminal work [96], explains that wealth accumulation in this form (or capitalism) is supported by a neo-liberal ideology, which encourages a free market economy with minimal intervention from the state. He goes on to state that neo-liberalism is a form of political economy, that strives to accomplish the only mission of restoring the class power of the global economic elite. In this form of political economy, the state is the facilitator of the free market and protects private property, according to Harvey. However, instead of acting as a balancing mechanism, the state acts only on behalf of a few segments of the society, i.e., the capitalist class. Using various case studies, the author shows that all nations that have moved towards neo-liberalism, show a sharp increase in social inequality and the redistribution of economic resources to the upper classes. Harvey also moves on to explain how the neo-liberal ideology has now infiltrated the public space through state approved think tanks and universities [97].

To maintain this status quo of accumulation of wealth towards the elite groups, even the media engages in producing a wildly uneven distribution of stories that favors the upper class, thereby supporting the neo-liberal ideology. This primarily happens since the media itself is controlled, directly or indirectly, by these elites. Chomsky in his book [99] notes that even most of the seemingly objective, factual media reports cater to the requirements of these upper class elites. The media focuses on its revenue channels, and hence, favors the upper class corporate entities since they supply the media with their required advertising revenues. Even when the media criticizes an elite, it is because it benefits another elite, eventually benefiting the media house. Habermas [90] argues on similar lines, to state that the mass media simply acts as a mouthpiece for elites belonging to the state and private organizations, which results in manipulating public opinion. Additionally, research shows that a small number of powerful corporations control the mass media [24], resulting in shaping of the public opinion towards their own interests.

Our work is motivated by these foundational studies on political economy analysis. We specifically focus on the political economy around key policies in India, and study some aspects of it in this work. Using the framework proposed in this thesis, we see how the corporations are interlinked with the state to form a power structure, and how these links evolve over time. Such structures might have both direct and indirect impact on policy-making as suggested by Harvey. On the media side, we study how the different news-sources are biased towards certain aspects, constituencies, and political parties. These biases often result in agenda setting and framing, resulting in manipulating the public opinion to disproportionately favor or oppose certain ideologies. We also see how different policies are justified by influential entities in mass media, social media, and the Parliament, to get an idea on the policy discourse in these forums.

3.1.2 Some Applications of Political Economy Analysis

Impact of policies on social welfare: Some works on political economy analysis specifically focus on the political economy of antipoverty and social safety policies in middle and

low income countries, and analyze the factors that contribute to their success or failure. For instance, Desai [64] studies the political economy around antipoverty programmes across various low and middle income countries, and analyses the reasons behind their failure, both in implementation and scaling up. Among them, he finds the pressure of countries to enforce trade-openness; lack of awareness and political clout of beneficiary groups; social and political fractionalization; and disinterest of policy makers in investing in pro-poor programmes, significantly important. Haggard et al. [91] study the political economy around social policies in the middle income countries of Latin America, East Asia, and Eastern Europe, and analyze the development and reform of the policies in these geographies, and the causes behind them. The authors suggest that the difference in realignments of governments, development strategies, and growth environment led to different paths of evolution for the social policies of these regions. Bossuroy and Couduel [40] study the political economy around social safety nets (SSNs) in various African countries, and suggest ways to expand and sustain them. The authors suggest several ways to to make SSNs successful, among which changing the political discourse around SSNs to avoid misconceptions, opportunity identification in changing political environments, and contribution by international platforms like the UN, are some. They also suggest various ways to enforce social accountability and obtain feedback of beneficiaries. These studies thus describe some applications of the methods of political economy analysis, and perform content analysis of social welfare and country level economic data.

Impact of policies on public goods distribution: Given the diversity of India in terms of its culture and political landscape, it is a rich source for the study of political economy. There exist several studies on various facets of political economy in the Indian context. For example, Banerjee et al. [27], study the influence of colonial power, landowner-peasant relations, and caste based divisions on availability of public goods in India. The authors show that in terms of access to public goods, the non-British controlled areas do much better than British controlled areas. On similar lines, Banerjee and Somanathan [28] study the political economy of public goods to understand how they are allocated across various constituencies in rural India, between 1970 and 1980. They find that among the disadvantaged social groups, those that could politically mobilize

themselves gained better access to public goods than others. The authors emphasize the importance of explicit political commitments related to infrastructure improvement and universal access to public goods during this period. Michael Levien [122] studies the political economy around land dispossession in India, in the context of *Special Economic Zones* (SEZs). Based on an 18 months study on an SEZ in Rajasthan, an Indian state, he establishes that land acquisition in its current form enhances class and caste inequalities in novel ways, marginalizes women, and creates an agrarian change that is disadvantageous for the rural farmers. Besley and Burgess [35] develop a model to measure government responsiveness in terms of food and calamity relief, and establish the importance of local language newspapers in transmitting information to the vulnerable citizens, thereby building pressure on the government to respond timely.

The studies described in this section focus on the applications of political economy analysis methods. While our study is motivated by these works, our goal is to build an information tool that helps users analyze the political economy around policies. The proposed tool uses computer based tools and techniques along with qualitative content analysis to generate insights on the political economy, which provides users understanding of the major actors in the policy process, the interconnections between them, and the policy narrative in the mass media, social media, and Parliament.

3.2 Frameworks for Political Economy Analysis

There exist analytical frameworks and guides for political economy analysis. For example, the analytical framework created by Moncrieffe and Luttrell [136] provides a tool to guide DFID in designing and conducting analyses of the political economy of specific sectors and policy arenas. Edelmann [69] provides an overview of a selection of approaches, frameworks and studies to analyze and manage political dynamics of sector reforms. Some works [128, 78] present a survey of several tools and frameworks for political economy analysis. These tools analyze the political economy at the country level, sector level, or

based on the problem at hand (problem driven tools). Acosta and Pettit [129] present a step by step guide to help development practitioners identify the important actors and institutions in the political economy around new policies, and their motivations behind policy formulation. We follow some of these guidelines to identify the key actors in the policy process: institutions and power structures; actors, agents, and their ideologies; and outcomes of policy implementation. We have developed the corporate-government KB to identify the institutions and power structure formed by influential entities. The key actors, agents, their ideologies, and policy outcomes are identified from mass media, social media, and parliamentary data.

3.3 Methods of Political Economy Analysis

There exist different methods for political economy analysis (PEA). Based on the analytical approach used, political economy analysis can be partitioned into (a) Content based analysis, and (b) Network based analysis. We discuss here some research works in both of these directions.

Content Analysis: Several works in the area of social science perform content analysis of data for political economy analysis. In the book *Using Content Analysis* [160], the author describes the different types of qualitative data analysis methods. *Content analysis* deals with analysis of documented information in the form of text, media, or even physical items. *Narrative analysis* is used to analyze content from different sources, such as interviews of respondents, observations from the field, or surveys. *Discourse analysis* is similar to narrative analysis, but it primarily focuses on analyzing the social context around the communication between the researcher and the respondent. *Grounded theory* analysis uses qualitative data to explain the cause behind a certain phenomenon. It studies a variety of similar cases in different settings, and uses the data to derive causal explanations.

Some studies use content analysis to study speeches given by entities influential in policy-making. An early work by Tetlock [211] involves analysis of speeches by US senators to

assess if they were isolationists – politicians opposed to aid other nations through policy interventions. Seider [178] similarly studies speeches given by top US business-persons across 11 industries to identify the similarities and differences in their stand, in terms of business requirements and public policy. In a recent work, Mumford et al. [138] study the foreign policy speeches by three British politicians in an automated fashion to see how the new labour code has incorporated an ethical dimension in the British foreign policy.

Content analysis has also been used extensively to study news data and policy documents. In her work [214], Elizabeth Thelen studies content of news articles or documents qualitatively to identify different features like valence or stance of articles towards an issue. Semetko et al. [179] use content analysis of news articles related to European politics to identify the dominant frames in them, and see how different news-sources vary from each other in terms of framing. Xu [230] studies the dominant frames using which the Occupy Wall Street protest movement was covered in the US media, using a mix of qualitative and quantitative analysis. Some researchers develop a coding schema to analyze the coverage distribution of topics on the concerned issues [195, 14]. There also have been qualitative studies on television news content to understand cultural and political trends. In this direction, Fields [74] provides a detailed methodology of qualitatively analyzing television news.

Some authors have performed content analysis on election data, data collected from government departments, public offices, industry associations, and corporations to understand the interplay between political and corporate entities, and to understand political behavior around elections [116, 118]. Huang et al. [103] perform content analysis on government documents related to environmental policies, to understand the priorities of the state with respect to various areas of environmental pollution in China. Verbeeten et al. [220] perform content analysis of CSR disclosure documents of 130 German companies over a four years duration, to find that CSR disclosures are positively correlated with firm value. Kamath [113] studies the extent of voluntary intellectual capital disclosure in Indian firms belonging to the information, communication and technology sector, and finds a significantly small extent of IC disclosures.

We use a mix of qualitative and automated content analysis of mass media, social media, and parliamentary data to understand some aspects of the political economy around policies, and to understand media bias in India. We next discuss in detail a few studies on content analysis of media data, which are close to our work in terms of the approach used. Elizabeth Thelen [214] studies reception of the transportation network company *Uber* in three different geographies by studying media articles published in these geographies, and performs qualitative content analysis to code the news articles into different aspects. The author also identifies the actors mentioned in these articles. Finally, the variation in coverage of these aspects and actors is used to explain the reception and reaction towards the service across the three geographies studied. In this thesis too, we study a few aspects of the political economy around policy events, which includes identification and coverage analysis of salient aspects (or topics) of these policies, and the entities that are most active in the policy discourse in Indian media. However, the difference between our work and that done in [214], is that alongside qualitative content analysis techniques, we also apply computer-based techniques that aid us in identifying aspects discussed in a policy event and the actors or entities discussing these events in an automated fashion from very large-scale web and media data.

Kaefer et al. [112] offer a step-by-step description of a software assisted qualitative data analysis of newspaper articles. They study 230 newspaper articles from the US, UK, and Australia on the international media's perceptions of New Zealand's environmental performance in connection with climate change and carbon emissions, using a multi-level coding approach. Based on this analysis, the paper depicts how New Zealand's earlier reputation has fallen significantly with respect to its environmental performance over time. Our work differs from this study in its analysis of a much larger scale of data from a wide range of topics in an automated manner.

Culpepper and Thelen [58] study the political influence of some popular technology firms like Amazon in their paper, and show that because of their huge consumer bases and reputation, they are at times able to bend policies in their favor. An example would be Amazon's resistance of taxation of Internet sales in California where the firm promoted

a referendum campaign against the legislation. The authors also discuss how articles on Internet privacy suddenly started seeing a rise in mainstream news-sources of several countries immediately after the Snowden revelations. As a source of potential influence of corporations on policy-making, we look into the alignment of viewpoints between business-houses and the state machinery on policy issues in mass media using automated sentiment analysis.

Costantino et al. [57] suggest a natural language processing (NLP) based technique to extract financial information from qualitative financial news data. They use different NLP tools to map each article to a template, which contains the most relevant financial information in a structured form. While the templates mentioned in this work are to some extent similar to the *aspects* that we automatically identify for each policy event in our work, we analyze news articles from a much larger spectrum of events rather than just analyzing a specific type of news data. To the best of our knowledge, this approach is unique and has not been attempted, especially in the Indian context.

We automate certain parts of content analysis in our work, like data collection, identification of aspects discussed about the policies, and sentiment and polarity analysis, and apply it to mass media, social media, and question-hour data. It is worth mentioning here that this work is methodologically similar to the work done by Thelen [214] and other papers discussed earlier, which use similar content analysis approaches [112, 58, 57].

Network Analysis: Researchers have adopted network analysis as an important method to understand political economy. Correia [56] examines firms and business-persons in the US with long-term political connections, and finds that they incur lower costs from the enforcement actions by the Securities and Exchange Commission (SEC), through PAC contributions and lobbying. Hai et al. [92] study the process of policy formulation in developing countries, and identify a set of relevant actors in the policy-making process. This set includes elected officials, appointed officials, business actors, labour, public, think-tanks and research organizations, political parties, mass media, and interest groups. Chen et al. [52] construct a political connection index to capture variations in the strength of corporate-political relations in China that incorporates various channels through which

a firm and its business-persons are politically or bureaucratically connected. Under network analysis, researchers have developed different frameworks to study policies. Shahaf et al. in their paper [186] propose a technique of connecting entities occurring in the news articles, thereby forming a network of entities, to help users navigate the news topic and add context to the news presented by media outlets. The GDELT Project [121] similarly monitors the broadcast, print, and web news globally in different languages and performs an analysis to identify the people, locations, organizations, themes, sources, emotions, quotes, and events, covered in the news. It combines content analysis with network analysis, and build networks that connect the different entities underlying sociopolitical events of interest.

Stasko et al. developed a framework named *Jigsaw* [199], which is similar to our system in terms of the representation of entities and the interconnections between them in the form of relationships. While Jigsaw solely uses text documents for discovery of entities, and captures the interconnections between them based on co-occurrence of these entities, our system uses both textual news articles, unstructured web sources, and structured data sources to identify these interconnections precisely. Moreover, our system is specifically focused on exploring corporate-government interlocks unlike Jigsaw, which works in a general domain. There are open platforms that provide corporate and political data in the form of network of entities.

OpenCorporates [207] shares data on corporate entities from many countries, and has been used to uncover several trans-national corporate ownership networks. Littlesis [105] shows connections between powerful people and organizations by tracking the key relationships of politicians, business-people, financiers, and their affiliated institutions. Unlike these platforms however, our goal is not just to collect and host such data but to build specific applications and provide inbuilt analysis tools which allow easy exploration of the data to obtain insights. Further, data about India on these platforms is not as rich as the dataset we have put together by integrating multiple data sources together. We build upon the aforementioned studies by looking at the corporate-political networks based on data from several web sources, and create a knowledge base (KB) of corporate-political entities.

This KB helps us capture the interactions and relationships between these entities, which eventually helps us demonstrate an approach to detect hot-spots of cronyism, and develop an empirical measure of the extent of interlocks between the corporate and government entities using an index of interlock.

3.4 Media's Impact on Political Behavior

Different forms of media have a significant impact on the political behavior of users. Here, we discuss studies that focus on the roles of mass media and social media in shaping political behavior.

3.4.1 Mass Media's Impact on Political Behavior

Mass media is known to have a significant impact on the political behavior of its readers as shown in several studies. Garramone and Atkin [81] develop a survey based method to find the correlation between print news reading and political participation (inclination to run for office, anticipated campaigns), and establish that print news has a significant impact on political behavior of youngsters. Gunther and Mughan [88] discuss the role of media in shaping the public's political attitudes by providing use cases from different geographies. There are other papers that show that even policymakers consider mass media to be an important source of political impact. Aelst and Walgrave [219] perform a comparative survey of the members of parliament in four democracies, and show that even the parliamentarians consider the mass media to be a prime political agenda setter, competing with the Prime Minister and the powerful political parties.

The significance of impact of mass media outlets on political behavior motivates us to study the discourse on Indian policies in mass media, and the entities that are covered by mass media in this discourse. However, unlike the aforementioned works, we do not directly analyze the impact of mass media on political behavior. Rather, our goal is

to understand the political economy around these policies; specifically, which aspects of these policies do the policymakers, bureaucrats, business-persons, and other prime entities cover, their sentiment with respect to these policies, and how they are interlocked. We also study how the audience of these media outlets express their opinions on the policies on social media.

3.4.2 Web and Social Media's Impact on Political Behavior

The web, as an information system, is known to significantly impact the political preferences of users. Some recent studies provide evidence of web based search engines impacting voting preferences of users through ranking bias, and other search features. Epstein et al. [71] demonstrated how ranking bias in web based search engines can lead to significant manipulation of voting preferences. This phenomenon is called Search Engine Manipulation Effect (SEME). Diakopoulos et al. [67] show how the editorial choices made by Google's algorithms shape information curation. The study proves that focus on official sources in rankings, ordering of different issues in the issue guide, dominance of a small set of news-sources presented in the *In the News* section, and differences in the visual framing can significantly impact voting choices. Hu et al. [102] study the search snippets produced by Google Search and the corresponding source web pages, and show that the snippets generally amplify partisanship when compared to the actual sources. Robertson et al. [172] develop a Chrome extension to survey users and collect the Search Engine Results Pages (SERPs) and autocomplete suggestions corresponding to political queries, and find significant differences in the composition and personalization of politically related SERPs, indicating differential impact of search engine results on on users' political information consumption.

Social media, also called the fifth estate, has now become an essential medium through which people produce, spread, and consume information. Manuel Castells [49] discusses how the advent of information and communication technologies (ICTs) have made social networks efficient, autonomous, organizational forms, and how social networks that use

these ICTs have become the basic units of modern society. Social media can act as a distribution tool to further distribute the information published in news-sources, and it can also act as a source of information working in parallel to the mass media. People often refer to social media posts, blogs, and other non-mainstream news outlets to get more complete information in addition to what the mainstream news-sources publish. For instance, Olteanu et al. [146] provide a comparative analysis of the aspects covered in mass media and in social media pertaining to events related to climate change, and find that there are significant differences between the two – social media provides a much higher coverage to independent climate activists, legal actions on climate change, and original investigative journalism compared to mass media. Social media have also been used to seek information about situational updates like circulating real-time information about disease outbreaks [206]. A Pew Research Center paper has shown that a significant number of adult Internet and social network users seek and post health related queries and information on social networks [79]. Social media can also act as a distribution tool for mainstream news-sources. Hermida et al. [100] show that a significant percentage of Canadian users use social networks like Facebook and Twitter to obtain news on current affairs.

Similar to the mass media, social media also has a positive effect of information diffusion leading to mobilization of citizens towards or against a political or social issue. In fact, today the social media has replaced mass media as the foremost platform that impacts political behavior of the users [150]. Gary King et al. [119] demonstrate in their paper how news media impacts people through social media exposure, and causes the American citizens to take public stands on policy related and national issues. Bond et al. [39] carry a controlled trial of political mobilization on 61 million users on Facebook during the 2010 US congressional elections, and observe that strong ties are instrumental in spreading real world and online behavior in social networks. Ang et al. [17] find that countries with sizable youth bulge, and more importantly access to Information and Communication Technologies (ICTs) in the form of social media have a higher prevalence of social or political protests. We do not study the impact of social media on political mobilization of users in this thesis. However, we do find evidence of increased awareness about political

issues among Twitter followers of mass media houses, in the form of sharing of URLs related to policy events.

The aforementioned studies motivate us to understand if the social media behaves any differently than the mass media in terms of the coverage and sentiment of policy aspects, i.e., if it paints a different picture than the mass media about the political economy around policies. Similar to works like that by Olteanu et al. [146] and Sutter [206], we study the coverage of policy aspects in social media, and see how it differs from that of mass media. This comparative analysis aids us in understanding if the social media acts as an information source parallel to and independent of online mass media.

3.5 Analysis of Bias

Since media impacts the political behavior of users significantly, it is essential for it to be balanced and free from any bias. However, this is not the case as has been observed in several studies. We cover some of these studies in this section.

3.5.1 Mass Media Bias

Given the impact that mass media has on the political behavior of people, bias of any form in the representation of policies in mass media has a detrimental impact on public opinion. Journalists and news-sources shape public opinion by intentionally or inadvertently creating bias in their selection, writing, and distribution of news content, and for this reason they have often been called *gatekeepers* [166]. Gatekeeping of information in turn leads to agenda setting, framing, and priming [176]. These three effects together play a significant role in influencing public opinion on sociopolitical issues. Biases present in online mass media also get amplified further due to readers' inherent biases that get reflected in their web browsing habits. This is studied in the US scenario by Flaxman et al. [76] where the authors find that most people visit a handful of ideologically similar

news outlets, leading to less diversity.

Different methods have been devised by researchers to study mass media bias. There are researches that study the impact of media bias on public opinion, by studying voting patterns of the audience of these media houses. For instance, Chiang et al. bring [53] out evidence of endorsements provided to political candidates by mass media in the USA. Some other works analyze mass media bias in terms of media's slant towards political parties, or media's coverage of policymakers. Gentzkow and Shapiro [82] developed an index to define a measure of media slant by analyzing key phrases in news content specific to political ideologies in the US. On similar lines, Munson et al. [169] assign a political bias score to each media outlet based on whether liberal or conservative candidates are over or under represented in these outlets. Some studies use crowd-sourcing to see if media outlets report in a partisan or non-partisan manner. Budak et al. [44] use crowd-sourcing and machine learning techniques to understand whether or not the US media reports in a non-partisan manner.

Our work is along similar lines where we use computational techniques with some level of manual fine-tuning, to build a structured method of analyzing alignment of news-sources towards specific aspects, constituencies, and political parties. While our automated method of clustering news articles into aspects can be used as a generic method to evaluate agenda setting, our manual mapping of these aspects to five constituencies or frames of perception can be used as a generic method to evaluate framing. Thus, our work can be considered to be a combination of the ideas that the aforementioned studies propose. Additionally, we go a step ahead and see how the bias in Indian mass media is propagated among its social media followers.

Mass media bias might occur in terms of the coverage of different aspects or entities pertaining to the policies (coverage or selection bias), and the way in which these aspects or entities are covered (statement bias). Using the research question: *Is the mass media biased in how it represents different policies?*, we try to see if the Indian mass media is biased in any of the aforementioned ways. In this direction, studies by Chiang et al. [53] and Ribeiro et al. [169] motivate us to see if the mass media is biased towards or against

any entity or a group of entities (like political party). Gentzkow and Shapiro [82] in their paper perform an analysis of media slant, which motivates us to understand the sentiment and coverage bias of mass media towards the key policy aspects.

Machine learning based models to study ideological bias: Significant research has also taken place in the area of detecting bias in terms of the political ideology of documents that include news-sources, blogs, political speeches, and political documents based on advanced machine learning techniques, especially *topic modeling*. Most of these technique build a model of bias, which can be used to predict ideological alignment of documents. Some studies in this direction study political documents like legislation bills to study bias in these documents.

Gerrish et al. [83] develop a machine learning model by combining *ideal point model* [54] with supervised topic modeling, which predicts voting patterns of legislators from the text present in the bills. Some other methods use political speeches to identify the ideological alignment of the speakers. Sim et al. [190] use an HMM based model to predict mixture of political ideological positions in political speech documents. Nguyen et al. [141] use hierarchical topic modeling to detect ideological bias in articles, product, and movie reviews. Recent work on deep neural network based approaches, like that of Iyyer et al. [107] focus on developing a recursive neural network based model to predict the political positioning of candidates from Congressional debates in the US. Unlike the aforementioned studies that are primarily based in the US, our work is based in the Indian context. We do not perform an analysis on the political ideologies of news-sources in our work, since we currently do not have ground truth data. Therefore, we have built a separate method to compare different newspapers on five dominant constituencies of *poor, middle class, corporate, informal sector and small trades,* and *government*, through which they present the information. These constituencies are the dominant *frames* of news presentation in Indian mass media. We use an automated approach to detect the dominant topics of discussion (or aspects) in the mass media on policy issues, and a qualitative content analytical approach to map these aspects to the five constituencies. Our method of extracting dominant frames or constituencies from the content presented in

mass media can be improved further, to detect ideological bias even in the aforementioned cases for web data and social media.

3.5.2 Web and Social Media Bias

Several studies have pointed out that social media often has its own biases, and that it also leads to misinformation and creation of echo-chambers. Biases also exist in the web and in the algorithms that recommend data to users. Here, we present the related work corresponding to each of these areas.

Web and social media bias: Search engines and online social media platforms have been argued to create biases in content distribution and display, typically initiated by inherent biases that exist among the users, which are amplified further algorithmically [21]. There are several studies that point specifically towards existing biases in content of web and social media. Garimella et al. in their paper [80] study the polarization of users on Twitter in terms of the content they post on controversial debates, and find that this polarization increases with the increase in interest about the event. News-sources and influential journalists can impact audience attention highly through social media as well, thereby propagating their inherent bias through social media [149]. In the absence of algorithms to support diverse information sharing however, users need to diversify their own networks to get a wider perspective as noticed on Twitter from the follower network of journalists and different media sources [15].

Self-selection bias: Social media also suffers from self-selection bias where the users form closely knit clusters owing to homophily based on similar information needs. Self-selection bias also occurs in web based search engines. Eli Pariser [155] discusses the issue of personalization on the Internet through Google's segregation of its user base into different filter bubbles. Similar biases have been noticed in the linking pattern of social media such as blogs [10] which leads to echo chambers. Such a view is corroborated by Quattrociochi et al. [164] where the Facebook news feed was argued to not be accentuating

the bias algorithmically than whatever bias already existed in the network as part of relationships defined by the people. Among these relationships, *weak ties* were found to be a better source of getting access to diverse information, as hypothesized by Granovetter [87] and also noticed in the works by Seth et al. [185] and Quattrociocchi et al. [164].

Algorithmic fixes to bias: Biases not only exist in the content produced by news outlets and social media, but also in web based search engines and their recommendation algorithms. There are studies that discuss ways to counter algorithmic biases of these algorithms. Celis et al.[50] study a variant of the traditional ranking problem in the presence of fairness or diversity constraints where they consider the value of placing an item in a certain position in a list (based on important attributes of the items), and a collection of fairness constraints to output a ranking, which maximizes this value while satisfying the constraints. On similar lines, Zehlike et al. [233] define and solve the *fair top-k ranking problem*, by developing a ranking algorithm to maximize utility, while maintaining group fairness constraints. Under these constraints, they ensure that for any position in the ranked list, a minimum proportion must be maintained for the group that is underrepresented. Our work is motivated from these aforementioned studies. We define the utility based on the exposure of *aspects* corresponding to a policy event in a news-feed. The utility varies across aspects for a policy event, and using a well defined utility function, we ensure that the aspect representation in our news-feed achieves fairness over long term and diversity over short term, while also ensuring recentness of articles displayed. A difference between the aforementioned studies and our approach is that we consider a temporally evolving set-up, and ensure fairness and diversity across multiple lists. On the contrary, the aforementioned works operate on a single list of items.

There also have been studies on building systems to counter the biases existing in online news data. Park et al. [156] develop a novel system named *NewsCube*, which automatically detects topics corresponding to events in popular news, and provides a balanced viewpoint to the user by ensuring plurality in the topics displayed. Munson et al. [139] similarly design and deploy a browser widget that aids in nudging its users to read balanced political viewpoints, alongside recording the aggregate political lean of users' read-

ing behaviors. Park et al. [157] design a framework to perform classification of news articles belonging to an event based on the different viewpoints from which the articles are written. This framework allows users to form their own, independent viewpoint based on a deep analysis and fine-tuning of the various aspects. While most of these studies are based in the US political scenario where the political leaning or bias can be bipolar, we identify biases in online news data based on multiple aspects, constituencies, and political parties in India, and ensure fairness and diversity of news presentation with respect to the aspects.

We study biases in social media (Twitter) content primarily as an extension of the analysis of mass media bias in this thesis. From the studies mentioned in this section, we see that biases exist in the social media outlets both in terms of the content hosted, and with respect to self-selection. These biases are intertwined, and also result in amplification of biases that already exist in information sources like the mass media. Our research question: *Are some news-sources more closely aligned with their readers on social media than others?* attempts to analyze these biases in social media platforms like Twitter, which are exhibited by the followers of news-sources. Detecting such biases are important, since they reflect the extent of independence that netizens show in terms of the content that they post. This aids in countering the biases existing in mass media sources.

We also described some of the researches done in the area of algorithmic fixes to biases existing in news data and search engines. These studies motivate us to answer the research question: *Can we produce a news-feed that is unbiased and fair, in terms of its representation of news?* This is a well explored question in the research community, and the aforementioned researches highlight the need to develop measures to counter the impact of bias in news data and recommendation algorithms in general. This thesis contributes towards development of a news-aggregator that aids us in answering this question in the Indian context.

3.6 Representation of Policies in the Parliament

Mass media and social media are the two major participants of any democracy. In the aforementioned sections, we focused on works that show how they influence public opinion, and how they might also be biased in their content. Another major stakeholder of democracy is the Parliament, which is the source of policy discussion and formulation. In this thesis, we also study how representative in its deliberation the Indian parliament is, i.e., if it provides equitable representation to the issues of all sections of people. Our work is motivated by studies that analyze the representativeness and deliberativeness of democracies. For instance, Gutmann and Thompson [89] define deliberative democracy as '*a form of government in which free and equal citizens (and their representatives), justify decisions in a process in which they give one another reasons that are mutually acceptable and generally accessible, with the aim of reaching conclusions that are binding in the present on all citizens but open to challenge in the future.*' Joshua Cohen [55] emphasizes that deliberative democracy emphasizes on involving public deliberation on different issues involving the citizens focused on the common good. The deliberative nature of democracy is considered important, since it actually considers the immediate and future concerns of all sections of citizens (and their representatives), through active deliberation on the feedback of the beneficiaries.

There have been several studies that analyze parliamentary data to understand the discourse around policies along different dimensions. Some of these studies analyze biases in parliamentary discourse with respect to variables like gender [37], some others look at the relative dominance that different political parties exhibit [20], while some others focus on analyzing political motivations of the parliamentarians (MPs) behind their parliamentary arguments [38, 173]. Bailer et al. [25] specifically analyze the representativeness of parliamentary discourse towards the concerns of different sections of people, alongside studying the motives of the parliamentarians in representing a certain cause. Unlike these studies, we do not delve deep into studying the motivations of policymakers to ask questions pertaining to policy issues. Rather, we try to understand how the questions asked in the Indian Parliament reflect the aspects discussed in the mass media, and whether

these questions discuss about the immediate issues of all sections of people. Researchers have also developed computational models to analyze parliamentary discourse. As discussed earlier, Iyyer et al. [107] use machine learning approaches to identify the political positioning of candidates from Congressional debates. Blidook and Kerby [38] build a regression model of parliamentary question-asking to demonstrate that Canadian MPs adapt their behavior in the Parliament to meet their electoral needs. We do not work on building a computational model specific to parliamentary data, but build a framework that aids us in doing a comparative analysis of the parliamentary policy discourse with that in the mass media and social media.

We explore a new direction in the study of parliamentary data where we see if the parliamentary discourse on policies in India is representative of the concerns of all sections of people. The corresponding research question that we try to answer is: *Is the policy-making process democratic, i.e., it ensures equitable representation of all sections of people and their problems?* This analysis, in turn, leads us to understand if the mass media and the social media are aligned with the Parliament in their representation of policy issues, or do they in any way vary from the Parliament, and work as independent sources of reporting about policies, with their own inherent biases in content.

3.7 Critical Perspectives of Big Data Analysis

In this thesis, we study mass media, social media, and other web data to analyze the political economy around key Indian policies. Our findings are thus based on quantitative and qualitative analysis of large-scale unstructured data. However, big data analysis has over time received several criticisms from multiple domains. Since this thesis primarily deals with collection and analysis of big data, we elaborate in this section some of the critical perspectives of big data analysis, and the ways we have tried to handle these issues in our work.

Criticisms around big data analysis occur at three levels. Some of these criticisms are

at the level of errors introduced by automated analysis tools [147]. The second level of criticism arises around the interpretation of patterns observed through big data analysis, even when the underlying tools analyzing the data are perfect [125]. A reliance on quantitative data without an understanding that emerges through qualitative data and mixed methods based research, can often be a cause of drawing incorrect interpretations. Finally, the third level of criticism occurs around how the data is obtained, especially when personal data of people is exploited for profit-making [171].

Although detailed research around these critical perspectives is beyond the scope of this thesis, we have tried to address each of these concerns in our work. We have performed validation of the accuracy of the analysis tools we have used or developed to prevent misinterpretations arising due to incorrect results or tool artifacts. For example, we have checked the accuracy of the sentiment slants output by the sentiment analysis tools (Sentistrength) after manually going through a large number of articles. Similarly, we also have checked the sanity of the aspects provided by LDA manually. To ensure that we do not misinterpret the patterns extracted by these tools, we performed detailed qualitative analysis validated through inter-coder reliability methods. For example, the aspect to constituency mapping exercise was preceded by the design of a detailed coding schema, which was finalized after much due deliberation. The exercise was also followed with measurement of the inter-coder reliability. Finally, all the data we have used is from public information sources such as newspapers, company registries, etc. None of this is associated with any personal information of people or that required informed consent from them to obtain the data.

Chapter 4

Analysis of Corporate-Government Interlocks

In this chapter, we take a look at the evolution of interlocks between the corporate and government networks in India over time as a potential source of influence in policy-making. The broad research question we try to answer here is: *How can corporate-government interlocks be identified that may have a potential influence on the policy process?* Corporate executives who enter politics or government administration, or family members of public officials with links to corporate organizations, are known to build an interlocking network that becomes a power structure of highly influential entities [133]. This power structure often leads to a bidirectional flow of favors between the corporations and political entities [72], and it can influence policy formulation or manifest itself in cronyism [161, 13], which can lead to financial benefits, positional privilege, or any other kind of power that leads to the formation of a closely knit class of influential people. With diminished competition and increased corruption, cronyism also impacts how well the state is able to redistribute wealth and create policies for social welfare [70]. Such negative manifestations of corporate-government interlocks have been evident in India [228, 101]. Cronyism causes, and has a positive feedback loop with inequality in income and wealth distribu-

tion. Given the growing wealth inequalities and fears of elite capture of public institutions to influence policy [202], it is therefore important to develop indicators and tools to monitor corporate-government interlocks. Having information about changes in the interlock structure can give researchers and journalists valuable context to interpret economic and policy changes. The data that we use in this analysis has already been described in the *Research Methodology* section.

Table 4.1 shows the details of the data collected for the knowledge base. In the next section, we describe the network computation done on the corporate-government KB to rank the entities based on the different patterns of interlock as discussed in chapter 2.

4.1 Related Work

Corporate-government interlocks are an indicator of strong collaboration between the state and corporate actors, which can lead to the formation of influential power structures. Mills critiques the network of power in the United States, which has significantly shaped the economy and government in the book *The Power Elite* [133]. On similar lines, Stiglitz talks about income inequality that results from these networks through rent-seeking and bidirectional flows of benefits between the corporate and political domains, and a positive feedback loop which sets in because increased inequality makes it easier for influential people to leverage their networks for personal gain [202]. There have been ample studies in this direction, we list here only some of them.

Flow of favor from the political to the corporate domain: Mian et al. [131] show in their work from Pakistan that bigger, politically connected firms have access to loans of much higher amounts from government banks as compared to unconnected firms. Another study by Mara Facio [72] carries out its analysis across several countries and observes sharp increases in stock prices of companies whenever they form political connections. Studies done by Fisman et al. [75] in Indonesia, and Johnson et al. [110] in Malaysia, similarly show that the fortune of politically connected firms is highly dependent

on the fortunes of the politicians they are connected to.

Flow of favor from the corporate to the political domain: Bertrand et al. [34] find that in France, in the wake of municipal elections, firms with politically connected CEOs see a sudden rise in employment rates (intended to glorify the concerned politician's image), and a drop in job destruction (firing). In the Indian context, Sandip Sukhtankar [204] provides evidence that during election years, politically connected sugar mills pay lower prices of sugarcane to farmers – this saving by the sugar mills is used to fund election campaigning, and is later passed on to farmers through waivers and other public policies if the politician wins.

Interlocks between bureaucrats and politicians/companies: While there are several works on the study of corporate-political interlocks, there also exists a wide range of work on interactions between bureaucrats and politicians. Interlocks between bureaucrats and politicians are mostly implicit, i.e., there may not exist any direct, explicit relationships. In such cases, the interlocks rather exist in the form of preferential appointment or transfers of bureaucrats by politicians. Several studies have argued that these appointments or transfers do not occur solely due to administrative needs, but because of personal favoritism of politicians, and flow of favors between politicians and bureaucrats.

Sanjoy Bagchi in his book [23] discusses cases where bureaucrats collude with politicians to obtain career benefits, and how politicians at times execute filters like caste to form a nexus of favorite bureaucrats. On similar lines, Jeffrey and Larche [108] discuss in their paper the rise of the Bahujan Samaj Party (BSP) in the Indian state of Uttar Pradesh in mid-1990's, when the party primarily seen as serving low caste and scheduled caste (SC) interests targeted several welfare programs designed for the low caste, and attempted to form a low caste hegemony in the Indian civil service through transfers of civil servants and placement of SC officers in key posts. Iyer and Mani [106] study micro level data on Indian bureaucrats (historical career data on IAS officers serving in October 2005) and politicians to prove that politicians use frequent reassignments or transfers to control bureaucrats, to serve their political interests, and provide evidence that caste affinity to a politician's party base aids bureaucrats in securing important and coveted positions.

Such political interference on bureaucracy also leads to problems in delivering public service transparently. For instance, Vaishnav et al. [218] and Kapur et al. [115] discuss how the Indian Administrative Service (IAS) is hamstrung by political interference, which leads to a detrimental effect on its performance in implementing public schemes effectively. Bibek Debroy [63] discusses the different measures that are required to protect Indian bureaucracy from undesirable political intervention. Most of these measures revolve around setting up an objective and transparent criteria for transfers, appointments, and promotions of bureaucrats. There also exist works that study flow of favors between bureaucrats and corporations, and reports signifying such interlocks. Barbara Harriss-White [94] studies the relationship between local state officials and merchants in the state of Tamil Nadu in India, and shows that merchants (business-persons) openly bribe state officials and take advantage of unofficial markets in licenses and public sector jobs. On similar lines, Paranjoy Guha Thakurta and Abir Dasgupta discuss in their article [213] how an IAS officer accused KPMG of exercising influence on government officials (bureaucrats) by recruiting their children and relatives. Other allegations include the offer of a bribe to a bureaucrat, and the awarding of crucial government contracts to international firms. Direct links between IAS officers and companies can also be observed abundantly. As reported in [188], several IAS officers quit the administrative service midway and join the private sector. We observe the same trend in our own dataset. Our corporate-government knowledge base provides evidence of several bureaucrat-company interlocks in the form of board memberships. Specifically, we see that several IAS officers join boards of large corporations post-retirement.

While the works discussed in this section provide evidence of specific cases where interlocks manifest themselves, there is no easy way or a service where researchers and journalists can analyze interlocks, identify curious patterns, and investigate these red-flags in more detail. Our corporate-government knowledge base (KB) and the platform to analyze this KB is an effort in this direction, to make it easier to identify undesirable outcomes from corporate-government interlocks. Specifically, our infrastructure allows us to answer the research question: *How can corporate-government interlocks be captured, which have a potential influence on the policy process?* Ours is a contribution in this area.

4.2 Details of Network Computation

We start with a global adjacency matrix (R), extracted from the social network (knowledge base) graph, which contains all the entities and their relationships. We then compute R^4 to obtain a matrix which incorporates interconnection information over 4 hops, and then obtain sub-matrices of R^4 such as a company-to-company matrix, or a politician-to-company matrix, to obtain mutual associations between specific types of entities only. These association matrices are used for the PageRank computation described below, along with also computing a set of bias vectors which capture entity specific weights independent of the pattern being considered[1]

The steps described below are for the pattern of *corporate connected bureaucrats*. Entity scores for other patterns are obtained similarly:

1. Apply PageRank on the company-to-company adjacency matrix (M_{cc}) obtained as a sub-matrix from R^4, with normalized authorization capital of the companies as the bias vector (E_c), to obtain a scoring of companies based on their corporate connectedness with other companies (C_n).

$$C_n = \lambda * M_{cc} * C_n + (1 - \lambda) * E_c$$

2. Multiply the bureaucrat-to-company adjacency matrix (M_{bc}) with C_n to get a scoring of bureaucrats based on their connectedness with companies (B_c).

$$B_c = M_{bc} * C_n$$

3. Obtain the hybrid scoring of bureaucrats (E_h) by linearly combining their corporate

[1]As an example, the set of properties used to build the bias vector for bureaucrats are selected by using OLS regression analysis based on a hypothesis that considers their membership in the board of a company as a dependent variable, and includes independent variables such as the number of weeks spent in foreign training, their educational qualification, number of weeks spent in important departments, their designation, appointment in central ministries, and total tenure till date.

based score vectors (B_c) with their bias vectors (E_b) containing information about their bureaucratic strength.

$$E_h = \beta * B_c + (1 - \beta) * E_b$$

4. Apply PageRank on the bureaucrat-to-bureaucrat adjacency matrix (M_{bb}), with the hybrid vector obtained in step 3 as a bias vector.

$$B_n = \lambda_1 * M_{bb} * B_n + (1 - \lambda_1) * E_h$$

Thus, the final scoring of bureaucrats (B_n) is made to depend on their connections to companies, their bureaucratic strength, the strength of companies to which they are connected, and their connection to other bureaucrats. We call this rank score of an entity with respect to a pattern as the **entity-score** of the entity. The basic structure of our set of PageRank equations are the same for other patterns as well.

4.3 Indicator Monitor Application

The goal of our interlock indicator monitoring application is to make it easier for users to observe changes in the corporate-government interlock over time, and discover curious patterns worthy of deeper investigation. To do this, we develop an indicator to quantify the strength of the corporate-government interlock, and enable closer investigation of the reasons behind changes in the indicator values over time. Our main finding is that the interlocks have strengthened over the years, most prominently through appointments of bureaucrats in the corporate sector after their retirement, and an increasing concentration in the corporate sector through denser company ownership networks and interlocks of shared directors especially across large companies. We define the interlocking network as the bridge edges that connect any entity on the government side (politicians or bureaucrats) with any entity on the corporate side (companies or business-persons).

This includes edges of five types namely, *politician—works in—company*, *politician—related to—manager*, *bureaucrat—works in—company*, *company—donates to—political party*, and *company—donates CSR amount to–politician*. The first three types of edges include *works in* edges, which indicate an entity's membership in the board of a company; and *related to* edge, that indicates family relationship between two entities. Thus, we capture both direct memberships in a company's board and indirect connections to a company through family links. The fourth edge type captures the relationship between a company and a political party through donations that the company makes to the party, primarily for electoral campaigns. The last edge type captures the amounts donated by companies under the corporate social responsibility (CSR) head. Under this type of edge, we consider a company connected to a politician if all of the following conditions are met: (a) the company donates a certain amount as CSR to a particular sector (e.g. health, education, etc.) in a constituency, (b) the Member of the Parliament (MP) elected from that constituency is also a minister, (c) the ministry governed by that MP also includes the sector for which the donation is made.

The indicator is computed as the sum of the scores of the bridge edges in the interlocking network:

$$I_{cp} = \sum_{i=1}^{|E|} edgescore(e_i)$$

where e_i is the i^{th} bridge edge – an edge connecting one node of the corporate network to another node of the government network. E is the set of all bridge edges. If u and v are the two nodes connected by a bridge edge, the score of the bridge edge is given by:

$$edgescore(u, v) = influence(u) * influence(v)$$

where *influence* is the normalized node score or normalized rank for a corporate (companies, directors) or government (politicians, bureaucrats) node as obtained from the PageRank based heuristic that we follow to rank the entities. We rank entities based on five patterns of corporate-government overlap namely, *corporate connected politicians,*

corporate connected bureaucrats, politically connected firms, politically connected directors, and *bureaucratically connected firms* [2]. We first validate our method of calculating this indicator, and next see how the interlocks between corporate and government entities evolve over time, and the reasons behind it.

4.4 Methodological Analysis

The intuition behind defining the indicator through the equations aforementioned comes from two basic principles:

1. The number of connections between the corporate and government entities is directly proportional to the extent of overlap between the two communities.
2. For a bridge edge, if one of the participating nodes has a low influence, while the other has a high influence, its overall score should be low. In other words, a bridge represents a strong tie only if both of its participating nodes are influential enough.

The following sections contribute towards mentioning some special cases with respect to this indicator, justifying the choice of normalization used, and finally, validating the indicator.

4.4.1 Special cases of entities captured by our ranking

This section describes some special cases that arise from the ranking process we follow for the entities. We capture the indicator for the four election years of 2004, 2009, 2014, and 2019, when the Parliament is inducted with new members who are elected for the corresponding term. The connectivity, and hence, the rank of these members or entities

[2]In case of calculation of the node score for a company, we consider the node score for it as the sum of the scores obtained by the patterns *politically connected firms* and *bureaucratically connected firms.*

change in each election year, since their connectivity with other entities increase over time. While the score assigned to each entity by our network captures the importance in the form of connectivity of these entity in most cases, we also observe that the change in rank of certain entities is not always highly correlated to the change in their scores. We list here a few cases of entities belonging in the government side, who have known connections with corporate entities, and are also captured by our network computation as some of the top ranked entities:

- **IAS officer Devi Dayal:** in the first slot (up to 2004), Devi Dayal was ranked 1331 among all IAS officers (for the pattern corporate-connected-bureaucrats). However, in the immediately next time slot, he ranked sixth. Given this huge jump in rank however, the change in probability score of this IAS officer is only around 0.0002 (from 0.0007 to 0.0009). It must be noted here that Devi Dayal as an IAS officer served in multiple influential ministries like Finance, Petroleum, and Agriculture. He became an executive director at Jindal Saw Ltd. in July 2004. He currently is connected to multiple other private organizations post retirement.

- **IAS officer Dileep Raj S Chaudhury:** in the third time slot (up to 2014), Dileep Raj S Chaudhury had a rank of 331. He jumped to the third position in the last time slot. In this case, the change in score has also been significant (from the order of 10^{-16} to 0.12). He is Director on the Board of IL&FS Water Ltd and IL&FS Paradip Refinery Water Ltd.

- **Politician Sanjay Dattatreya Kakade:** this politician jumped from a rank of 6601 to 7 in the last two time slots (for the pattern corporate connected politicians). In this case too, the change in score was significant (from the order of 10^{-8} to 0.00014). He is a BJP MP and a real estate developer (Kakade Infrastructure and several other firms).

We see such cases for all of the patterns considered. In an earlier study, we also found that there is a significant intersection between the top 10 percentile highest ranked entities in

each pattern, and the top 10 percentile highest connected entities. In other words, we find that the rise in ranks of several entities is directly correlated with the increase in their number of connections[3]. Thus, our network computation appropriately captures the importance of entities based on their corporate-government interlocks. We use the ranks assigned to the entities by our computation to calculate the indicator of interlock.

4.4.2 Normalization of node scores

In our case, the influence of an entity node can be quantified in two ways: (a) by its node score, and (b) by its rank. The number of nodes for each type of entity is very different in our case. For example, we have close to 20K politician nodes, around 70K company nodes, and 7K bureaucrat nodes in our dataset. Thus, while calculating the bridge scores, the influence for a node must be normalized within each type of entity. We considered several normalization techniques among which the following two were experimented with:

1. **Minmax normalization**: Since our network is very sparse for the time slot of up to 2004, for each type of node, the node scores follow a nearly uniform distribution for this slot. On the other hand, towards the later years, we see a sharp skew in the distributions, where a few top nodes get very high scores, while the rest of the nodes form a long tail of very low node scores. Thus, the minimum and maximum scores for the same set of entities vary widely for two consecutive time slots. To handle this issue, we performed minmax normalization of the node scores before calculating the indicator (so that across all time slots, the scores are normalized between 0 and 1).

2. **Rank normalization:** Since the number of entities for each type is very different in our case, the rank of two different types of nodes cannot be compared directly.

[3]The new connections added over time are mostly the ones that form later – for example, a bureaucrat forming a new connection with a firm as its director. In cases where the connection is an existing one, but is captured later by our data collection process, we ensure that the connection is considered at all earlier timestamps when it existed. For instance, if a family relationship is captured later in our data collection process, we augment it to the set of connections considered for all previous timestamps.

For example, the rank-5th politician node might not have the same influence as the rank-5th company node, as the number of company nodes is much higher than the number of politician nodes. Hence, we followed rank normalization to normalize all of the ranks (for each type of entity) between 0 and 1.

Each type of normalization has its own pros and cons. We finally considered rank normalization to be the best in our scenario. We elaborate further on this in the following sections.

4.4.3 Comparing the indicator with a random baseline

The only way to validate our indicator from real world data is to consider articles and reports related to corporate-government connections from the mass media or other relevant data sources. However, there is no easy way to aggregate the inferences from millions of these documents to see if our findings align with the big picture presented by these sources. Another way is to refer to experts on the topic, to see if their opinions match our findings. While we are currently working in this direction, the exercise might suffer from difference in opinions and subjectivity.

Hence, we considered comparing our indicator with a random baseline. To create a random scenario, we randomize our corporate-government network multiple times, and calculate the indicator of overlap for each randomization iteration. We call the indicator of overlap calculated on our random network I_{rand}. For a specific time slot, we see the distribution of I_{rand} and its median value to see if the median lies below our actual indicator I_{CP} (henceforth, we call this median value across all randomization iterations for a specific time slot I_{rand}). If $I_{CP} > I_{rand}$ for a time slot, we can state that I_{CP} appropriately captures the extent of overlap, and that it is not by a random chance that the connections (bridges) were triggered. Hence, we verify the following hypothesis:

Hypothesis: *The indicator of overlap I_{CP} captures the overlap between corporate and government entities, which is not a random, chance event (i.e., the bridge edges are*

formed by the entities with certain intentions, and do not occur by random chance).

We initially considered randomizing the entire network, while maintaining the degree of every node, and the type of nodes that they connect to (so that edges are not randomly formed between incompatible node types). However, this did not seem logical, since it would lead to change of the intra-government and intra-corporate connections as well. For example, in this randomization approach, not only would any politician connect with any company, but the politician could connect to any random political party (as a member) or any random department as well. This would lead to change of the intra-network properties of the politician (properties within his political network). Our goal is to retain the intra-network properties (political parties of politicians, departments where the bureaucrats have worked, and so on), while randomizing only the inter-network (corporate-government) connections, i.e., the bridge edges. In the next section, we report our experimental results on this set up, where we randomize only the bridge edges.

4.4.4 Randomizing only bridges: minmax normalization

We ran 60 randomization iterations on the network. For each iteration, we calculated the indicator based on the node scores normalized by minmax normalization. The actual indicator I_{CP} was also obtained using this normalization for this experiment. We find that in this case, I_{CP} lies above I_{rand} in two of the four slots. Hence, we the hypothesis is not fully satisfied for all time slots using this scheme.

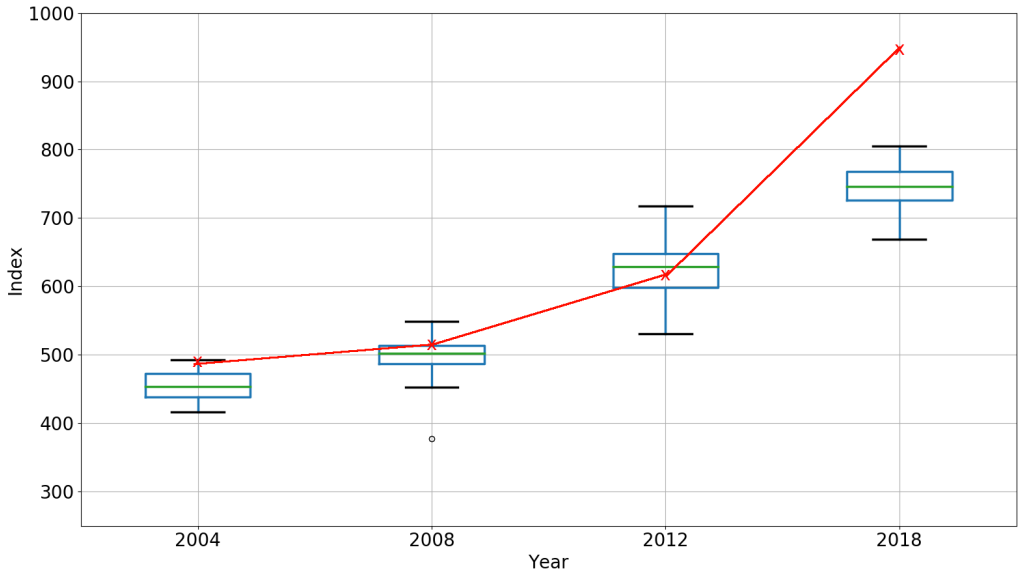
Why not minmax normalization: Minmax normalization maps the probability distribution of scores of nodes between 0 (minimum) to 1 (maximum). Hence, it basically distributes our original node scores over a bigger range (since for the original distribution $min(score) >= 0$ and $max(score) <= 1$). Post minmax normalization, even in the [0,1] score range, relative difference between the node scores remain as they were. Thus, it does not really address the issue of the large gaps in node scores for closely ranked entities (or the lack of correlation between node scores and node ranks). The problem with

these huge gaps of closely ranked entities is that they bias the weights of the bridges. For example, if entities A and B are two consecutively ranked entities, which form bridges with the same node X, intuitively, bridge A-X and bridge B-X should not vary much in their weights. Since there is a huge difference between the node scores of A and B, bridge A-X might now weigh incomparably higher than bridge B-X, biasing the indicator.

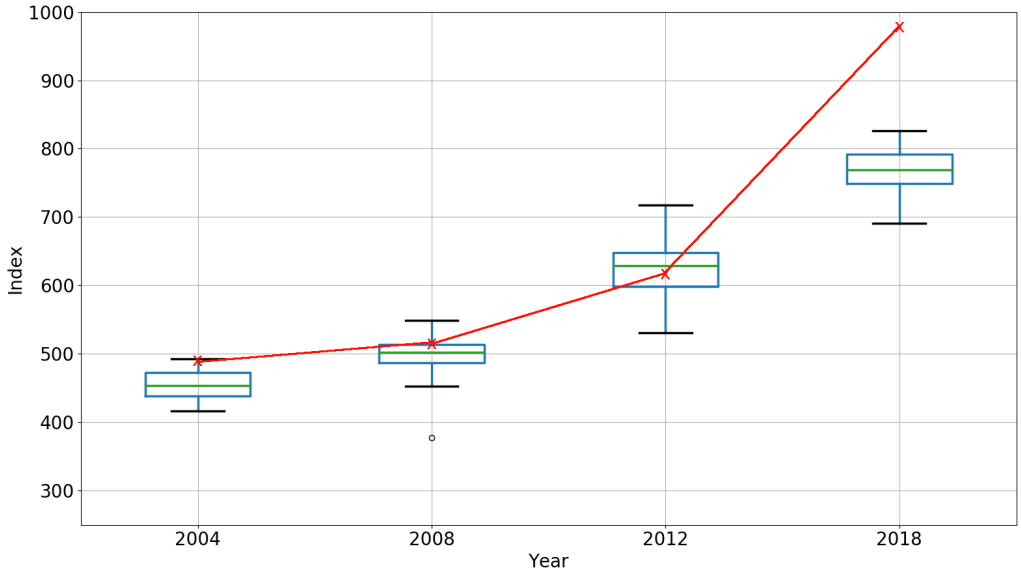
Another problem of minmax normalization arises because we have widely different number of nodes for various entity types. In such cases, applying minmax normalization results in assigning high rank scores to each entity belonging to the entity group with smaller strength. For example, we have 20K politician nodes and 90K company nodes in our dataset. Applying minmax results in assigning a higher rank score to each politician, but a lower rank score to each company. But we should do the opposite, since a top ranking entity in a bigger entity group should have a higher influence (rank score) than a top ranking entity in a smaller entity group. This is another reason for not applying minmax normalization in calculation of our indicator.

4.4.5 Randomizing only bridges: rank normalization

Owing to the aforementioned reason in the previous section, we consider rank normalization, which normalizes the ranks of the nodes between 0 and 1. For the indicator calculation, we now use these normalized ranks (instead of the node scores). Note that apart from the advantage rank normalization provides (described previously), it also alleviates the problem of large gaps in node scores. We ran 60 randomization iterations on our network as before in this case, and apply rank normalization to calculate both I_{CP} and I_{rand}. As discussed earlier, we have implicit and explicit links in our dataset. Here we show two plots – one where we have considered only the explicit links to calculate the indicator, and the other where we use both kinds of links for the calculation. We find that our graph now partly verifies the hypothesis in three out of four time slots in figure 4.1 for both explicit and implicit+explicit indicator calculation. Only for the third time slot, we see that the original indicator is slightly less than the median of the random



(a) Only explicit links



(b) Implicit+explicit links

Figure 4.1: Indicator plot for the four years with rank normalization: the blue curve denotes I_{CP}, and the box plots denote I_{rand}

indicator. However, even in this case, we see a significant percentile of I_{rand} values lying below I_{CP}. It must also be noted that the median random indicator of overlap, I_{rand} also increases over time similar to I_{CP}. This is because the corporate-government network is continually expanding as new people and new relationships are included over time, and the old links or people are not excluded from the network generally. This trend can be observed from table 4.2 as well. We thus observe that the overlapping bridges between our corporate and government networks are not formed by random chance, and that there might be specific interests in play, leading to formation of connections between influential actors.

4.4.6 Data limitations

For the first time slot (up to 2004), our data is sparse, and there are many connections that have not been captured during this time period. This subjects our computations during this time period to some level of doubt. Apart from this, our data also suffers from the following problems:

- We have few politician to company connections, primarily because these connections are often not present in the publicly available websites from which we have crawled the data. Many of these connections are also in the form of shareholding, which we have not captured. Mass media data was also considered as a source of collecting relations between politicians and companies, but we were not able to find many relations between the existing entities. We have introduced CSR links and corporate donations to political parties in our data to find out indirect and implicit links between the politicians and companies, which have been included in our analysis.

- Cliquishness of data is another problem, especially in the politician network. Since the politicians can belong to only one political party at a time, they are generally interconnected only to the other politicians in the same party. Due to lack of data, it is difficult to break the cliques formed in this manner, due to which our PageRank

based network computation suffers. We already have included family member based relationships to alleviate this problem. However, the data still needs to be enriched to get newer links that break the cliquishness of the political network. We have used mass media data to extract more family relations for this purpose. Incorporating correlated movements (especially in the form of politician-department-bureaucrat links) can also aid in this.

- Due to some errors in entity resolution, some of the connections incorporated in our database are prone to error. Currently, we discovered that 34 connections in our database are erroneous due to faulty ER. We are in the process of removing and correcting these connections.
- Due to unavailability of data, we have very few timestamped edges. In case of bureaucrat to company connections (where the data is richer), although we have information about the start dates of these links, the end dates are unavailable in many cases. We will work towards filling this gap as part of our future work.

We see that the indicator of corporate-government overlap shows trends that indicate towards constantly increasing corporate-government interconnections. Provided that we enrich our data further, the methods developed to compute this community overlap could be applied to get newer and more accurate insights.

4.5 What are the causes behind increase in the interlocks?

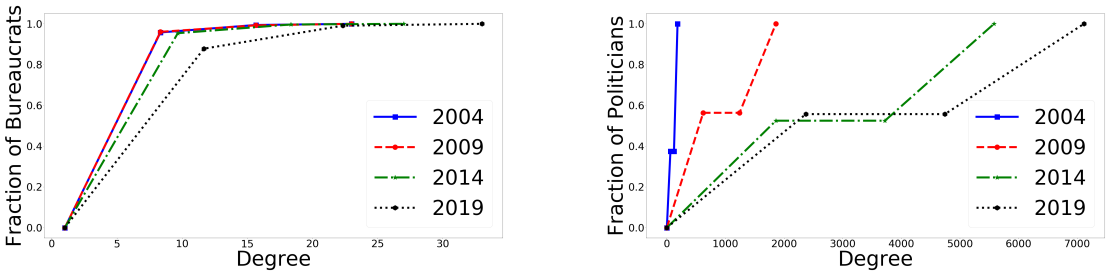
There can be two structural reasons behind the increase observed in the indicator values: (a) the formation of new bridges over time, and (b) increase in scores of the existing bridges due to either an increase in concentration in the government network or in the corporate network. To investigate this further, in table 4.2, we show how the number of bridge edges for each timestamp.

Bridges	Untimed	Before 2004	2004-2008	2009-2013	2014-2018
POL-COM	58	14	12	14	10
POL-BoD	72	27	33	32	36
IAS-COM	296	39	132	264	357
IAS-COM (Govt./Public)	654	5	77	381	834
COM-PP (donation)	–	125	216	978	1172
CSR	–	–	–	–	167

Table 4.2: Count of bridge edges added during each time period (untimed edges were considered for calculations across all time periods). POL, COM, BoD, IAS stand for politicians, companies, directors, and bureaucrats respectively. The Govt/Public links are for appointments of bureaucrats in state owned companies, and are not considered in the calculations.

We can see that the number of bridge edges corresponding to especially the *corporate connected bureaucrats* pattern significantly increases with time. This observation is validated by sources [231], which mention how bureaucrats are increasingly taking up corporate positions after their retirement. To study the second reason of an increase in concentration of the government and corporate networks, figure 4.2 shows the CDF of the degrees of the interlocking bureaucrats and politicians within each time period. There is a clear trend of increasing concentration from 2004 to 2018, showing that not only are new links being formed rapidly between the government and corporate sectors, but the degree centralities of the interlocking nodes are increasing as well, leading to an increase in the value of the indicator. These findings are consistent across all other types of entities.

We further investigate the interlocking network to check if it is becoming denser itself. The clustering coefficient (CC) of the interlocking network does not increase over time (Figure 4.3), showing that the bridge edges are being formed between different pairs of entities. However, the CC of the 1-hop, 2-hop, and 3-hop neighborhood of the interlocking network extended only on the corporate side shows an increase over the years, indicative of increasing connectivity within the corporate network, either through subsidiary links or shared members in the BoDs of companies. A similar investigation of the neighborhood



(a) Bureaucrats (b) Politicians

Figure 4.2: CDFs of degree centralities of interlocking bureaucrats and politicians

of the interlocking network extended on the government side however does not show any increase in the CC, validating that the increasing interlocks are happening due to increased concentration in the corporate network.

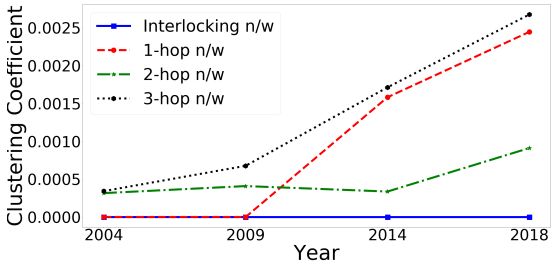


Figure 4.3: Change in clustering coefficient of the corporate-government network with time

4.6 Discussion and Conclusion

The research question that we answered in this section is *How can corporate-government interlocks be identified that may have a potential influence on the policy process?* We have developed a country level indicator to capture the corporate-government interlock, but our methods can be applied within a specific policy area to understand the political economy in which it operates. In this chapter, we presented a system that can be used to extract useful knowledge from web and media data to monitor the extent of corporate-government interlocks in India, which are a source of bias in policy-making and formulation. Our key findings from this analysis are as follows: (a) There is a monotonic increase in the extent of corporate-government interlocks over time, in the 2004-2018 period (b) This increase occurs primarily due to the increase in degree centralities of the interlocking nodes over time in the knowledge base graph (c) Although the density of the interlocking network does not increase over time, the densities of the 1/2/3 hop neighborhood of the interlocking network increases significantly on the corporate side

We thus find that the interlocking politicians, companies, and bureaucrats show a tendency to increase their interlocks over time. We have provided examples of IAS officers who have headed several important designations, and show a propensity to increase their corporate interlocks further. In a separate work, we also observe that these interlocks are mostly formed with larger corporations (in terms of authorization capital). This indicates towards formation of a power structure where highly placed government officials collaborate with large corporations. On the other hand, we also see corporate-political interlocks (both implicit and explicit) increasing over time, indicative of increasing collusion between policymakers and corporations.

While in this thesis, we do not focus on how this nexus between corporations, policymakers, and bureaucrats might lead to rent-seeking practices, it is an interesting direction to study, and our corporate-government knowledge base (KB) can aid in it. Our primary contribution lies in the collection, resolution, and storage of this large-scale data from multiple sources at one place in the form of a KB, and in the creation of the indicator

of overlap that aids in capturing the extent of corporate-government interlocks. There is a strong need to study corporate-government interlocks in a systematic manner. Our work contributes towards this by helping users track changes in the interlocks over time, thereby helping them understand how the corporations have a direct or indirect influence on policy-making.

Entity Type	**Count**	**Dataset**	**Attributes**	**Timestamp**	**Sources**
Politicians	19295	LokSabha	Name, constituencies, political party	2004-09, 2009-14, 2014-till date	indiavotes.com
		Rajya Sabha	Name, constituencies, political party	2014-till date	wikipedia.org
		Prime/chief ministers	Name, constituencies, political party	All till date	wikipedia.org
Bureaucrats	11531	IAS Officers	DoB, department, districts or locations served, educational qual., training details	1961-till date	persmin.gov.in
Business-persons	111105	Directors (managers)	Director ID (DIN), name	1961-till date	www.mca.gov.in
Firms	64155	Listed firms	Income, subsidiaries	1980-till date	bseindia.com
		Listed/non-listed firms	Company ID (CIN), name, headquarters, authorization capital, date of incorporation	1980-till date	mca.gov.in
Departments	1565		Name, ID	All till date	wikipedia.org
States	36	Name	Current		wikipedia.org
Family information	71	For politician	Name	All-till date	wikipedia.org
	222	For managers	Name		wikipedia.org
CSR donations	167	CSR donations	Amount donated, year of donation	2014-2017	csr.gov.in
Political party donations	2491	Donations to political parties by companies	Donation year, amount of donation	2009-2017	myneta.info

Table 4.1: Overview of web data collected for the corporate-government knowledge base

Chapter 5

Analysis of Bias in Mass Media Content

Content authored by news-sources and published on the web, is a significant source of web content and reflects the biases of mass media. We examine the coverage given on the Indian mass media and social media to the four economic policy issues as described earlier – Demonetization, Aadhaar, GST, and Farmers' Protests – to understand the existence of different forms of media bias. The media module extracts articles on each of these policy events using an augmented keyword based approach as discussed in the Data Collection section (chapter 2). We next run LDA on these articles to obtain 16 aspects for *Demonetization*, 14 aspects for *Farmers' Protest*, 11 aspects for *GST*, and 17 aspects for *Aadhaar*, and then name each aspect manually.

Our work is significant in creating a standardized methodology to assess the following: the ideological positioning of a news-source with respect to the various constituencies of people, and its alignment with the social media discourse of its community. The analysis presented in this section helps us answer the broad research question: *Is mass media biased in how does it represent different policies?* Under this research question, we answer the following sub-questions: (RQ-1) Are news-sources biased on the amount of

coverage they give to different aspects about the policy issues? (RQ-2) Do news-sources have a bias towards or against the five constituencies or frames of presentation? (RQ-3) Is the mass media biased towards one of the two major political parties? and (RQ-4) Are some news-sources more closely aligned with their readers on social media than others? In this chapter, we answer the research questions asked above, and present the relevant results.

5.1 Related Work

In this section, we position our work with respect to the extant literature in the field of media bias analysis and describe some of the studies that motivate the research questions that we ask in this chapter. We start with the studies on various techniques to capture mass media bias, and move on to how social media fails to counter this bias, by instead amplifying the biases in mass media, creating echo chambers, and spreading misinformation.

There are several ways in which mass media can be biased. Some media houses are biased towards certain aspects or issues of a policy, while some others are biased with respect to the political parties or entities (like corporations, bureaucrats, etc.) that they cover. Bias of any form in the mass media has a detrimental effect on public opinion – biased media produces a biased understanding of an issue. Hence, we are interested in understanding if the Indian online mass media houses also carry some of these biases.

Political bias of mass media: There are several works that study the political bias of mass media, i.e., biases in mass media with respect to political parties or candidates. Ribeiro et al. [169] use the Audience API provided by Facebook to study the audience demographics of the social network, and study the bias in the media sources followed by them on Facebook by scoring the online media houses based on their left or right leaning. While this work does capture the audience demography of Facebook based news-sources, it is difficult to apply this approach to mainstream news-sources as information about

their audience group is difficult to obtain. Budak et al. [44] study some of the major US news outlets and identify their political articles, to measure the political slant of these news outlets through crowd-sourced analysis, and find that all of the major newspapers report about political events in a non-partisan manner, and that political leaning can be identified by the disproportionate amount of criticism that the outlets subject a political party or candidate to. The problem of crowd-sourced analysis is that it suffers from subjectivity, and may not be able to capture the accurate picture besides being time consuming and costly. D'Alessio and Allen [60] study partisan media bias in presidential election campaigns since 1948. The authors consider gatekeeping bias, coverage bias, and statement bias of mass media outlets, news-magazines, and TV news towards political parties in the US. While no significant biases were found for the newspaper industry, their meta-analysis of studies of television news showed small and measurable coverage and statement biases.

Unlike the aforementioned studies, our method of identifying the entities (like political candidates, business-persons, and bureaucrats) covered in the policies is fully automated, since the scale of data that we work upon is large. Provided that a large number of works investigate political bias of mass media in different geographies, in the Indian context, we were interested in answering the research question: *Is the mass media biased towards one of the two major political parties?* While the studies described in this section are all done in the US context, we try to understand how the Indian mass media is biased towards the two major political parties (the currently ruling Bhartiya Janata Party (BJP) and Indian National Congress (INC).

Mass media Bias with respect to aspects or frames covered: Some other studies on mass media bias focus on the biases in mass media with respect to the different aspects discussed about events, or the frames through which these aspects are presented. For instance, Smith et al. [193] study the protests held in Washington D.C. in 1982 and 1991 and see if the coverage of protests in print and television news are framed in ways consistent to the views of the protesters. They find that the aspects covered vary between print and TV news, and so do the frames of presentation. While TV news is more

thematic and cover the issues that led to the protests, print news covers the incidents of interest (like violence) during the protests and tend to ignore the background issues. Another study by Boykoff [41] explores the framing practices used by mainstream mass media outlets in the US in their coverage of the Global Justice Movement during two major episodes: the World Trade Organization protests in Seattle in 1999, and the World Bank/IMF protests in Washington, DC in 2000. The author performs a content analysis of prominent news-sources along five dominant frames of presentation, and shows how different news-sources vary in the way they present the protests in terms of the aspects covered and the frames through which they are reported. Researchers have also analyzed mass media bias with respect to issues that are not chiefly political. Smith and Wakefield [194] perform a content analysis of tobacco related editorials from 310 US based news-sources. They study the aspects and frames of news presentation on the issue, and find that organizing frames like supporting policy interventions, condemning the industry, highlighting individual rights, and expressing general cynicism were most prevalent.

While our study of mass media bias in terms of the aspects covered within policy issues and their frames of presentation is similar to these studies, we study a much larger scale of data, and the time frame of our analysis is also much longer (2011 to 2018). Another difference lies in our approach of aspect and frame identification, which is a combination of computational and qualitative analysis that requires minimal manual intervention. While most studies in this space of mass media bias perform aspect analysis and frame analysis in an interleaved fashion, we attempt to segregate the two fields. To understand the coverage bias of Indian mass media towards certain aspects of policy issues, we attempt to answer the question: *Are news-sources biased on the amount of coverage they give to different aspects about the policy issues?* These aspects are then mapped to five constituencies or frames of presentation in this work, which leads us to our second research question: *Do news-sources have a bias towards or against constituencies like the poor, middle class, corporate, informal sector, and government?* These analyses aid us in placing the prominent news-sources on a 5-dimensional constituency space, which also helps us understand the difference between these news-sources with respect to their dominant frames of presentation.

The two broad types of content analysis as identified by researchers are *thematic analysis* and *relational analysis*. While the first approach analyses the primary themes present in the content based on frequency of certain keywords occurring, the second approach focuses on identification of themes along with understanding the relations between them. Our work can be classified under the first approach of thematic analysis where we analyze the dominant themes or *frames* of presentation of content in mass media. A frame is the way in which a content is presented to the audience such that it is processed in the desired form. For our data, we define five frames of presentation of news regarding policy issues: poor, middle class, corporate, informal sector, and government. Thus, a news piece containing keywords related to the poor is classified under the *poor* frame, and we next see the alignment of the news towards this frame (pro/anti poor).

The contribution of our work is that it helps us in automated mapping of the news presented in Indian mass media houses to a set of topics or aspects, which can in turn be mapped qualitatively to the set of five dominant frames of news presentation. Our system uses automated clustering techniques like *Latent Dirichlet Allocation* to identify the aspects the news article is about. These aspects provide us an idea of the dominant topics that are being discussed in the article, thereby helping us analyze their alignment towards or against the predefined frames. To the best of our knowledge, it is the first work that studies media bias by content analysis through a framing lens at a large scale. Our technique of identifying the aspects and frames of content presentation is also generic, and can be applied to the bias analysis of any news-source. It is important to mention that although we have applied our method to quantify media bias, but it can also help in detecting and analyzing echo chambers existent in social media, and also in spotting of fake news spread by them.

While biases exist in mass media in terms of policy representation, the audience also obtains information on current affairs from social media outlets. We have already discussed studies that show how the social media is used as an independent information source, in the Related Work section. Given that a significant number of people use social media even for news consumption, one may expect social media to counter some of the biases

that mass media exhibits. However, studies have established that social media as an information source is not always independent of the mass media, and does not counter the biases existing in mainstream media. Social media is also known to exhibit self-selection biases that lead to the formation of echo chambers, which leads to further biasing of the users' information requirements. Next, we describe a few works on the mass media's impact on social media, and social media users' tendency to form echo chambers.

Mass media's impact on social media: Researches have shown that not only does the social media at times not counter the biases carried by mass media, but that these biases are further amplified by it. Saez-Trumper et al. study 80 international news outlets, and find that they show selection and coverage biases depending on their geography of operation. They also observe that these biases are amplified by social media followers of these news-sources. Our work is in a similar direction. However, we use an automated clustering approach (LDA) to identify the aspects that are covered for a policy issue, unlike [174] where TF-IDF and cosine similarity based measures are used to identify *news stories*. Additionally, we perform our analysis of social media coverage of mass media articles on a per-policy basis where we see the trends corresponding to each policy.

Due to these biases and their impact on users, social media is also proven to be involved in agenda setting as a distribution channel for mass media sources. For instance, Feezell [73] employs a longitudinal survey based experiment conducted over seventy-five days in the spring of 2014, and see how the users' perception of importance of aspects change owing to their exposure to news snippets. The author proves that agenda setting on social media occurs through incidental exposure to political information among the least politically interested netizens. While we too study agenda setting on social media, we do it in an automated fashion by mapping the mass media aspects to social media posts of users based on the URLs shared. Our study is also done over a longer period of time on a per-policy basis. Mass media is also seen to impact platforms apart from traditional social media outlets, such as citizen blogs.

Sharon Meraz [130] studies traditional media blogs and independent citizen blogs in the US by analyzing hyperlink usage in blog posts. They find that although not the sole

driving force, the agenda setting effect of traditional mass media still plays a significant role in shaping the agenda of even the independent blog-spaces. We similarly study the effect of news-media on social media consumption and sharing of news by looking at the mass media URLs which the followers of news-sources share. However, our study is solely based on these users' social media posts, and not on independent blogs.

News-sources and influential journalists can often impact audience attention highly through social media, thereby propagating their inherent biases [149]. In this paper, the authors use a corpus of 1M tweets from 200 journalist Twitter accounts and audience responses to these tweets, and develop predictive models to identify the features of both journalists and news tweets that influence audience attention. While this work focuses on identifying the features that maximize audience outreach on social media, we study the coverage that mass media aspects of a given policy issue get on Twitter. Specifically, we observe the similarity that exists between the news coverage in mass media and social media, in terms of the aspect coverage and sentiment. The aforementioned studies on the impact of mass media bias on social media motivates us to answer the research question: *Are some news-sources more closely aligned with their readers (on social media) than others?* This analysis helps us understand whether the followers of Indian news-sources counter the biases in these news-sources through their independent social media posts (on Twitter), which in turn, clarifies if the agenda setting and framing effect of mass media has a significant influence on the social media space as well.

Echo chambers in social media: The influence of mass media's agenda setting effect on social media also leads to creation of echo chambers, which results in a skewed understanding and unbalanced opinion of different political and social issues. This effect is also corroborated by Bennett and Iyengar [31], who observe that technology has actually started to narrow users' political horizons, and most of the exposure to political news is voluntary and concordant with the users' existing political beliefs. Habermas also points towards this decreasing realm of the public sphere with the Internet taking its place [90], which leads to reduced discussion on diverse viewpoints.

There are several studies on community biases or self-selection biases in social media.

Some of these studies focus on the effect of content production and link formation based on content produced in social media that lead to echo chambers. For example, An et al. [16] find that the news sharing on social media sites like Facebook is partisan in the US context, and most people only share news that they find aligning to their own political beliefs. Ideological biases have also been noticed in the linking pattern of social media such as blogs [10] leading to echo chambers. Thus, these studies show that not only do users share content aligned to their own ideological beliefs, but they also show a tendency to connect to circles that reinforce them. Some researches show that these self-selection biases are generally user created, and do not generate from algorithmic biases existent in social media. For instance, Quattrociocchi et al. [164] observe that the Facebook news feed is not accentuating any bias algorithmically other than the existing biases in the friends' network on the social media as part of relationships defined by the people. Finally, some studies focus on the content consumption and interaction on social media. Bakshy et al. [26] examine how 10.1 million US Facebook users interact with socially shared news, and quantify the extent to which individuals encounter and interact with diverse content on Facebook's News Feed. They find that homophily plays a significant role in news reading behavior of users, and there are clear ideological biases in the way the social media algorithm suggests news to its users.

While these studies specifically focus on general social media users' sharing, consumption, or interaction with diverse information to see if they form echo chambers, in our work, we study social media data corresponding to the followers of news-sources that we study. We see the extent of overlap among the follower communities of news-sources, i.e., if the readers of news-sources try to form an independent and balanced opinion on policy issues by following diverse news outlets with varying political alignments on Twitter.

The aforementioned studies motivate us to see if echo chambers exist among the follower communities of Indian news-sources on social media. This direction of analysis also raises question on the causality of mass media bias: do the news-sources produce biased content owing to their followers' ideological positioning towards various constituencies, or are they inherently biased despite their followers being open and balanced? We describe

these analyses later.

5.2 Aspect Coverage Bias of Mass Media

The research question: *Are news-sources biased on the amount of coverage they give to different aspects about the policy issues?* relates to the *agenda setting* effect of mass media as reported in literature [176]. Agenda setting is the idea that there exists a strong correlation between the emphasis placed by mass media on certain issues, and the importance attributed by the readers to these issues. In this research question, we study the emphasis that mass media places on various aspects in terms of the relative coverage given to these aspects. We define relative aspect coverage for each news-source as follows:

$$relative\_aspect\_coverage = \frac{w\_a}{\sum_a w\_a}$$

where w_a is the number of words appearing in articles belonging to aspect a for a particular news source. In other words, for a news-source, we define relative aspect coverage as the proportion of words in the aspect with respect to the total number of words across all aspects in that news-source. Thus, we normalize the count of words per aspect, which handles the case of different news-sources containing different length of articles (and aspects).

When we look at the top aspects covered by mass media corresponding to each policy event in table 3, we find that for Aadhaar, the highest covered aspects generally relate to the middle class (example: *Aadhaar enrollment centers* that talks about logistics of Aadhaar enrollment, and *Court cases related to Aadhaar* that primarily covers data leakage and privacy issues) or to the effect of the policy on economy (*Positive effect on economy*). Demonetization as a policy involved significant political debates and the highest covered aspects are all political in nature (example: *Appreciation by PM and Opposition's statements against Demonetization*) followed again by an aspect related to the economy (*Negative impact of Demonetization on economy*). GST, which is an economic

policy issue, has the highest coverage provided to two aspects namely, *Political push to include petroleum products under GST*, and *Non-conclusive discussion between centre and states*. The first aspect here involved a lot of political discussion, but is also relevant to the middle class. The second aspect is a politically polar topic of discussion, which includes discussions on the losses incurred by the states in case GST is implemented. We thus observe that bias does exist in coverage provided to different aspects for each policy event, and the highest covered aspects are generally political or relevant to the middle class. However, although there is some discussion on the immediate issues of the traders and companies (*Confusions regarding GST rate slabs* and *Fears of capital crunch among traders*), the problems of the consumers – like rise in prices of essential commodities and services due to GST implementation – is not given significant attention. Finally, in case of Farmers' Protest, we find that the top covered aspects mostly focus on quick remedies to the problems of farmers (*Disbursement of loans* and *Loan waiver implementation by state governments*). There also is some discussion on structural issues related to the farmers' distress like *Irrigation concerns and water pollution*. The details of this analysis can be found in our paper that we have recently submitted.

We observe this trend even when we take a look at the highest covered aspects for each event, on a per-news-source basis. For a policy, the top five aspects covered by every news-source remain more or less consistent, and generally belong to the political or middle class domain. However, some news-sources like TeleG show a higher coverage for aspects relevant to the poor, when compared to others. IE, on the other hand, shows a higher skew (difference in coverage between the highest and lowest covered aspects) than other news-sources.

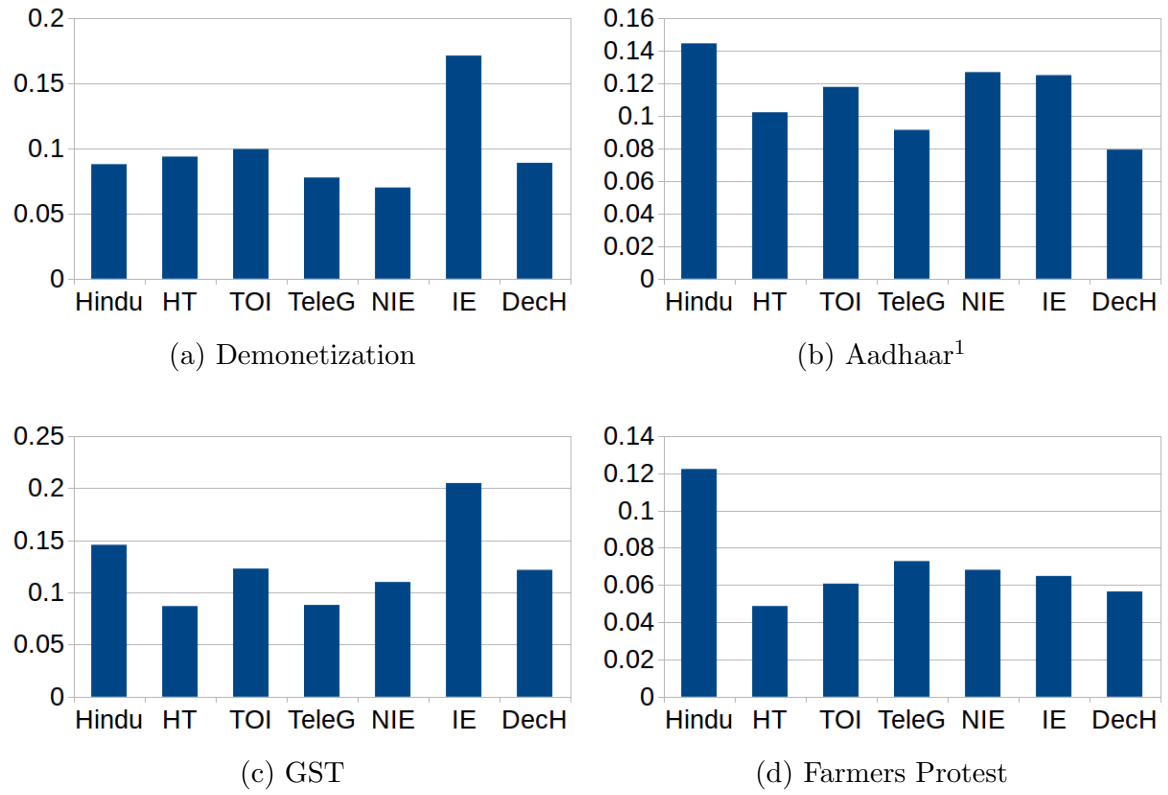
To see if the differences that seem to exist in the relative coverage given to various aspects by the news-sources are significant, we find the global or mean relative aspect coverage for each aspect, by averaging the relative aspect coverage across all news-sources. To then see which news-sources deviate the most from the global relative aspect coverage, we plot the euclidean distance between the two distributions (relative aspect coverage distribution for a news source, and the mean relative aspect coverage distribution across

Aspects (Aadhaar)	**Mass Media**	**TweetFol**
Aadhar enrollment centers	9.9%	8.9%
Court cases related to Aadhaar	9.9%	13.5%
International Linkages/Positive Effect on Economy	8.8%	13.6%
Implementation of Direct Benefit Transfer Scheme	6.9%	7.6%
Parliamentary debates on Aadhaar	6.3%	7.5%
Aspects (Demonetization)	**Mass Media**	**TweetFol**
Appreciation by PM for supporting Demonetization	15.9%	26%
Opposition unites against government on Demonetization	14.2%	12.6%
Negative impact of demonetization on Economy	8.3%	10.2%
Probes and Arrests of black money hoarders	6.1%	6.8%
Long queues at banks and ATMs and cash crunch	5.9%	4.1%
Aspects (GST)	**Mass Media**	**TweetFol**
Protest to include petroleum products under GST	16%	16.9%
Non-conclusive discussion between centre and state	15.2%	33.1%
Discussion in the Parliament in support of GST	13.4%	3.9%
Revolts/Confusion with GST Rate Slabs	12.8%	18.9%
Fears of capital crunch among traders before GST rollout	12.3%	4.4%
Aspects (Farmers' Protest	**Mass Media**	**TweetFol**
Disbursement of Loans and Subsidy by Banks for farmers	12.2 %	10.8 %
Irrigation concerns and water pollution affecting farming	9.7 %	22.8 %
Loan waiver implementation by State Govts	7.5 %	3%
Protests by farmers	7.1%	6.1%
Farmers' Distress regarding Minimum Support Price for Crops	5.6%	2.8%

Table 5.1: Relative aspect coverage for mass media and social media, for the top five highest covered aspects in mass media

news-sources) for each news-source in figure 5.1.

We find from these plots that the highest deviation from mean aspect coverage is mostly shown by IE and Hindu, both of which are commonly believed pro-opposition news-sources [134]. We also find that apart from IE and Hindu, most news-sources lie close to the mean aspect coverage trend (except in Aadhaar). This reflects that for most policies,



(a) Demonetization (b) Aadhaar[1]

(c) GST (d) Farmers Protest

Figure 5.1: [RQ2] Euclidean distance of relative coverage and mean relative coverage (across news-sources) for the four policy events. Higher the deviation for a particular news source, more different is its coverage from the mean behavior across news-sources.

most news-sources preferred to stay close to the average trend of aspect coverage in mass media. We also find the Kolmogorov-Smirnov 2-sample statistic for each news-source for each event, to see how significant the difference is between the news-source's coverage of aspects and the mean relative coverage. We find that the p-values lie in the range of [0.58,0.99] for Demonetization, [0.19,0.99] for Aadhaar, [0.74,0.99] for GST, and [0.26,0.86] for Farmers' Protest. Thus, we conclude that the difference between the relative coverage of aspects and the mean relative coverage is insignificant for all news-sources, indicating that nearly all news-sources follow the global trend of aspect coverage. The least p-value

is always seen for IE across all events, indicative of its non-conformist coverage of aspects.

5.3 Constituency Coverage Bias of Mass Media

Through the research question: *Do news-sources have a bias towards or against constituencies like the poor, middle class, corporate, informal sector, and government?* we try to analyze the effect of *framing* in mass media [176]. Framing refers to the modes of presentation that media houses use to present information in a way that aligns with the readers' underlying schema of perceiving the content. One of the ways in which news-sources engage in framing is by orienting the news content towards specific constituencies that their audience use to perceive the content. We analyze this effect by automatically extracting aspects from the news articles, and manually linking them with one or more of these five constituencies based on the coding schemes designed for each policy event. As mentioned earlier, this mapping simply tells us if articles in that aspect contain keywords semantically similar to the constituency name, or if the articles discuss about issues pertinent to the constituency. Next, we find the alignment of the aspect towards the constituencies in terms of the sentiment of its constituent articles. This is done in two steps: (a) we first find out the majority sentiment slant m of the articles in an aspect a, and (b) we then see if the majority articles of the aspect indicates whether it supports or is against a particular constituency c, i.e., its stance with respect to the constituency or *stance(a,c)*. We divide *stance(a,c) by the majority sentiment m, to get the alignment score (U) of the aspect with respect to the constituency. U* can vary between -1, 0, and +1 for each (aspect, constituency) pair. These scores are presented in the Appendix. Using these (aspect,constituency) alignment matrices for the four events, we calculate the (news-source,constituency) alignment matrix M as follows:

$$C_{ian} = \frac{c_i}{\sum_{j \epsilon (n,a)} c_j} \tag{5.1}$$

$$S_{an} = \sum_{i \epsilon (n,a)} C_{ian} * (S_{ian} - S_{avg}(a)) \tag{5.2}$$

$$M(n,c) = \sum_{a \epsilon c} U[a,c] * S_{an} \tag{5.3}$$

where n represents a news-source, a an aspect, C_{ian} is the relative coverage for the ith article, in news-source n, belonging to aspect a, and S_{ian} is the compound sentiment score of the ith article for aspect a. $S_{avg}(a)$ is the average sentiment score of all articles in aspect a across all news-sources, and c is the constituency. Here, $(S_{ian} - S_{avg}(a))$ is the offset of the sentiment of article-i from the mean sentiment of all articles for aspect a across news-sources, $U[a,c]$ is the (aspect,constituency) alignment value $\epsilon[-1,0,+1]$. Thus, the matrix M tells us how aligned a news-source is to the aggregate behavior of all news-sources, towards a constituency, in terms of the coverage and sentiment deviation with which it presents its content. To empirically verify if there exists variations (or similarities) in terms of constituency alignment of the news-sources, we performed a Principal Component Analysis (PCA) on the 5-dimensional mean constituency vector (mean of the 5-dimensional constituency vectors across all events) for each news-source, for the four events. Figure 5.2 shows the plot. A factor analysis was then done to interpret the two principal components we obtained here.

For the first component PC1 (x-axis), we find that the constituencies *informal sector*, *middle class* and *poor* are negatively correlated with the *government* and the *corporate* constituencies. This component represents if the news-source is aligned towards the informal sector, middle class, and poor (towards right), or towards the *government* or *corporate* constituencies (towards left). The second component represents alignment towards the *government*, *corporate*, and *informal sector* constituencies to the negative side.

We observe that TeleG, a commonly believed leftist news-source, is most aligned to the constituency *poor*, the *informal sector*, and the *middle class*. On the other hand, TOI is seen to be an outlier, and in general covers all of the constituencies much differently than the other outlets. DecH, IE, HT, and NIE, being close to the origin, are balanced

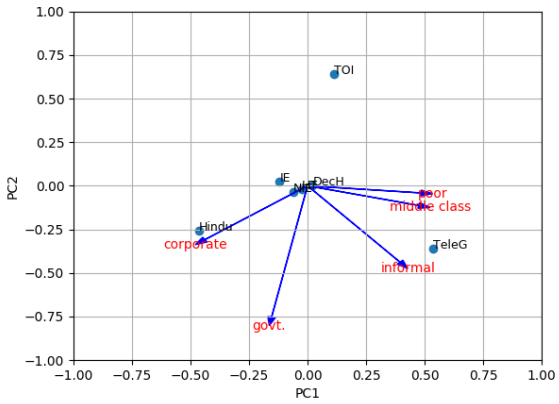


Figure 5.2: PCA on constituency vectors for the four events: Principal component PC1 represents news-sources that cover more of informal sector, poor, and middle class (towards right) related issues, and political or corporate related issues (towards left). Principal component PC2 represents news-sources that cover political, corporate, and informal sector related issues (on the negative side).

news-sources. Finally, Hindu is aligned more towards corporate and political discussions, and again covers these constituencies much differently than the majority news-sources.

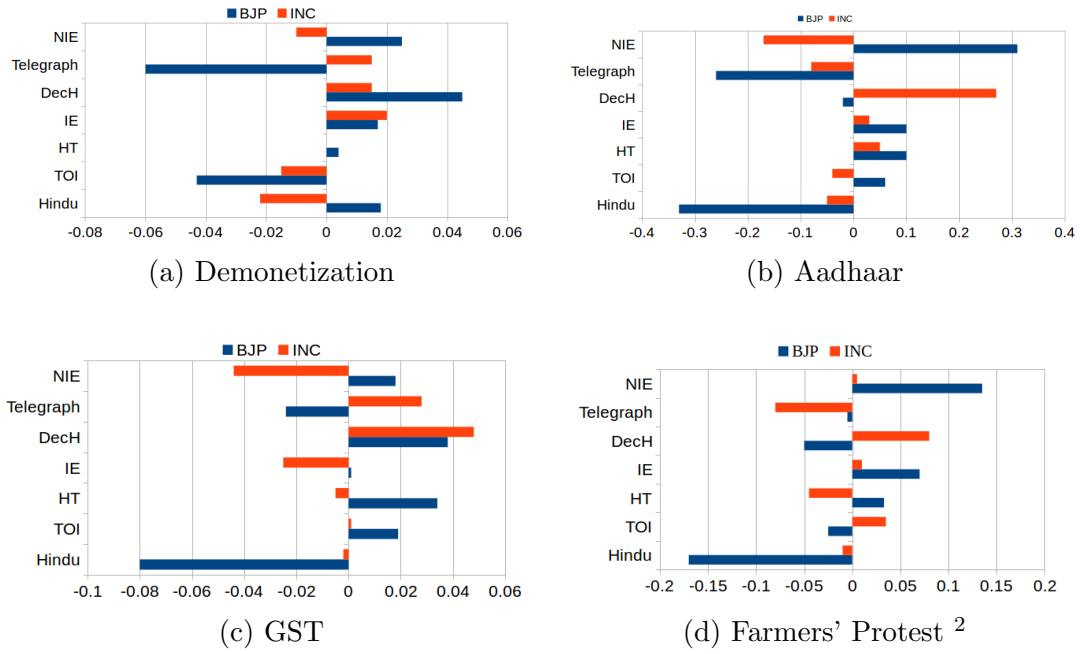
In a separate analysis, we compare the mean relative coverage provided to constituencies by mass media (across all news-sources), and find that the coverage is consistently higher for the *middle class* (above 50% coverage for Demonetization, Aadhaar, and GST) and *government* (above 90% coverage) constituencies, when compared to the *poor* (less than 50% coverage for Demonetization, Aadhaar, and GST). Thus, in terms of coverage of issues, we find that mass media in general provides less coverage to issues related to the poor, and more coverage to middle class and political issues. We present these findings in the Appendix.

The PCA analysis and the analysis of the mean relative coverage provided to constituencies by mass media indicates the presence of *framing* effects of mass media: (a) the news-sources are biased with respect to the five constituencies, in terms of the coverage

and sentiment with which they present their content, and (b) they provide consistently less coverage to issues of the poor in general.

5.4 Political Party Bias of Mass Media

In this section, we answer the research question: *Is the mass media biased towards one of the two major political parties?*. We see in figure 5.3, how much coverage is given to the two major political parties in India. BJP is the currently ruling party, while INC is the primary opposition. We consider the top 100 entities with highest coverage and manually



(a) Demonetization (b) Aadhaar

(c) GST (d) Farmers' Protest [2]

Figure 5.3: Deviation of relative coverage of entity groups from their mean relative coverage across news-sources. Mean coverage is taken as the average coverage of an entity group across all news-sources.

group them to their respective bins of political parties for this analysis. Next, we measure

the deviation of the relative coverage given to an entity group (party) by a news source from the global relative coverage across all the news-sources. Here we use statements *about* the entities, to see the sentiment in which the news-sources project these entities.

From our analysis, we find that our data shows nearly the same alignment of the news-sources as is the public perception [134]. Across events, NIE (commonly believed to be pro-BJP) has given above average coverage to BJP, and Telegraph (commonly believed to be leftist) has provided below average coverage to it. For INC, we find DecH and IE (both believed to be pro-INC) giving above average coverage, and NIE giving below average coverage to it. Finally, Hindu (believed to be a leftist news-source) seems to be non-partisan, in the sense that it mostly provides below average coverage to both parties. Our analysis thus reveals that some news-sources indeed show a bias in the coverage they give to different political parties, which also reflect their long-term or short-term political affiliations based on ownership networks of media. We plan to investigate this more closely in future work with an analysis of policy steps undertaken under different political regimes.

5.5 Alignment With Social Media Content

We answer the research question: *Are some news-sources more closely aligned with their readers on social media than others?* in this section. We analyze if the readers' preferences in terms of the importance placed on certain aspects, and the sentiment slants of their posts, correlate with that of mass media. We consider the readership community of news-sources as the set of all followers of the news-source handles on Twitter (*TweetFol*).

Under this research question, we see if the aspects tweeted by the readers of news-sources align with the ones presented by mass media. For each news-source, we compute the Jensen-Shannon Divergence (JSD) between the distribution of its aspect coverage and that of its social media community. The Jensen-Shannon divergence is a principled divergence measure that quantifies how distinguishable two or more distributions are from each other.

There are previous studies that have proven the effectiveness of JSD to measure distance between two probability distributions [59]. Similar to our study, Beretta et al. [32] used JSD to evaluate the divergence between gene expression profiles, and showed that the JSD values of profile pairs lying close to zero indicate similar gene profiles. Table 5.2 depicts our findings. We see that the news-sources have a high alignment[3]

News Source	Demonetization	Aadhaar	GST	Farmers Protest
	TweetFol	TweetFol	TweetFol	TweetFol
Hindu	0.12	0.08	0.17	0.15
HT	0.13	0.03	0.18	0.07
IE	0.14	0.03	0.28	0.08
NIE	0.11	0.04	0.13	0.07
TeleG	–	0.11	–	0.07
TOI	0.11	0.04	0.12	0.04
DecH	0.12	0.10	0.15	0.11

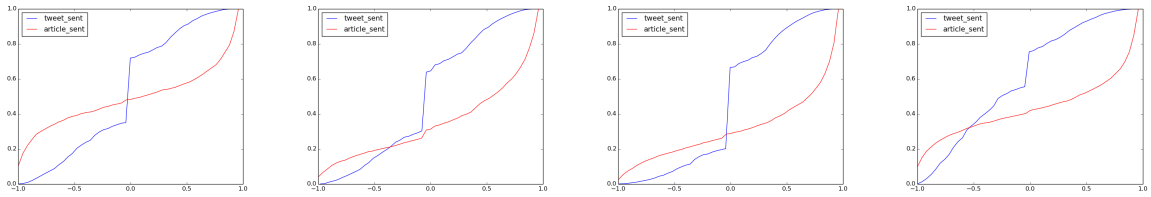
Table 5.2: [RQ3] JS divergence showing difference in aspect coverage between mass media and social media: for TeleG, we could not find any tweet for Demonetization and GST. The Kolmogorov-Smirnov 2-sample test also suggest that the aspect coverage are significantly similar between the mass media and social media.

We also perform the Kolmogorov-Smirnov 2-sample test for each event, for mass media and social media coverage of aspects, to see if the differences in coverage are significant. We find that the p-values lie in the range of [0.30,0.89] for Demonetization, [0.19,0.99] for Aadhaar, [0.37,0.99] for GST, and [0.11,0.86] for Farmers' Protest. For a sample size of seven, the D-values lie beyond the permissible range even in cases when the p-value is less than 0.2. Thus, we conclude that the difference between the distributions of mass media and social media coverage are insignificant. We are able to see hence that both the

[3]JS divergence is bounded in the range [0,1]. Lower the value of divergence, closer it is to zero, and higher is the similarity in terms of aspect coverage with their followers on social media (as seen from the low values of JS divergence), and the readers of the news-sources prefer to closely follow the aspect coverage trend of their favorite media houses. [32]

mass media and social media provide significantly similar coverage to the aspects under each policy event, thereby both providing give less coverage to issues of the poor, strongly indicating a bias arising in the web content produced in the mass media and social media. Our observation of social media followers of English news-sources providing significantly less coverage to the immediate issues of the poor (similar to what their favorite news-sources do) shows that the social media community, which chiefly consists of the middle class, is less keen on talking about the poor. On the other hand, the poor often do not have access to online social media to represent their own issues. This trend can be considered to be an effect of *Digital Divide*, defined as an uneven distribution in the access to, use of, or impact of information and communication technologies (ICT) between any number of distinct groups, which may be defined based on social, geographical, or geopolitical criteria, or otherwise [143, 222].

We next analyze for each news source, how much the overall sentiment of the Twitter posts on its news articles align with the sentiment of the original article. In figure 5.4, we show the CDFs for article sentiment and tweet sentiment (note that in this case, we consider all tweets by the followers of a news-source, irrespective of whether they contain URLs or not).



(a) Demonetization (b) Aadhaar (c) GST (d) Farmers' Protest

Figure 5.4: CDF plot of article sentiment and tweet sentiment for the set TweetFol, for *The Hindu*. For the other news-sources for all events, we present the results in the Appendix.

The plots indicate two observations: (a) Above 20% of the tweets by followers of news

handles are neutral (in Demonetization, Aadhaar, and GST more than 40% of the tweets are neutral), and (b) The article sentiment is either more positive or more negative than the tweet sentiment (as observed from the curves to the right and left of the neutral axis). Both of these trends indicate that the Twitter followers of news-source handles are more neutral than the news-sources themselves[4]. These findings tell us that social media seems to respond to whatever is being discussed in the mass media, but the sentiments of the social media readership are not entirely aligned with the news-sources that they follow.

Considering the high degree of alignment of aspect coverage between the mass media houses and their Twitter community, we next want to check whether these communities are distinct from each other. We therefore analyze the community overlap between the followers of different news-sources and present our results in table 5.3.

	TOI	HT	Hindu	IE	NIE	TGI	DH
TOI							
HT	0.57						
HINDU	0.44	2.59					
IE	0.79	3.71	6.21				
NIE	0.30	1.13	2.32	3.82			
TGI	0.46	1.68	3.06	6.08	10.38		
DH	0.35	1.07	1.99	3.49	16.25	41.68	

Table 5.3: Odds-ratio of overlap of follower community for each pair of news-sources

We broadly find that communities of DecH, TeleG, NIE, IE, and Hindu form a closely knit cluster in terms of their community overlap. Among these news-sources, DecH, TeleG, and NIE show the highest overlap of communities among themselves (odds ratio for community overlap > 10). HT and TOI have least overlap with others, and form outliers

[4]The sentiment slant is calculated using different methods for news articles and tweets – Sentistrength is used for the articles, while Vader is used for the tweets. However, we perform all of our analysis using the distributions of sentiment slant. So, this is not a problem.

(odds ratio for community overlap < 4). This indicates that many of the followers prefer to follow news-sources with commonly believed affiliations that are opposite to each other (DecH is commonly believed to be pro-opposition and NIE is commonly believed to be pro-ruling party). This might be an indication of users' tendency to consume news from news-sources with opposing polarities, to counter the information bias in media. Our findings partly support the findings from the study by Mullainathan et al. [137], which models the mass media to see the effect of competition among media houses on two types of biases – ideological bias (bias towards or against a political ideology), and spin bias (bias occurring due to propensity of media towards creating a memorable story). They observe that although competition can aid in removing ideological biases of media, it exaggerates the incentive of spinning stories. Our analysis provides a direction to study the open question of whether the newspapers tune their coverage to bring them closer to what their follower communities want, or whether communities of people gravitate to the news-sources that align with their ideologies, or both. It raises the question of what other factors come into play for social media followers to decide which news-sources to follow? This decision function seems to not only be dependent on the biases in the content, but may have more factors included, such as differences in the popularity of the newspapers, and preconceived preferences of the users. This is an open question, and can be investigated further to understand the rationale behind the choice of which news-sources to follow.

5.6 Discussion and Conclusion

The research question that we tried to answer in this section was *Is mass media biased in how it represents different policies?* We find that variation exists in the coverage of different aspects across news-sources, indicating media bias. Our analysis on mean coverage provided to each constituency also suggests that the news-sources generally provide high coverage to political issues and issues related to the middle class. On the other hand, the issues of the poor do not get enough attention in comparison. The

PCA analysis based on coverage and sentiment of the content also indicate biases in the alignment towards the five constituencies. We also find that social media is more balanced in taking up the views of academics and activists for discussion and distribution; politicians still get the most coverage but less than that given by mass media. However, the aspects covered by the readers of news-sources is closely aligned with the those covered by the news-sources themselves. This tells us that the biases existing in mainstream media is echoed by their social media followers, and that the followers do not offset this effect.

Comparing the findings of the research papers that we cited in this chapter, and our own findings, we see that our work validates the different theories of mass media bias and the way social media aids in furthering these biases. We find that the Indian mass media is biased in terms of the aspects that it covers with respect to policy issues, and the frames through which they present them. These findings are in line with studies like [193, 41] that show how the media is biased with respect to the aspects covered and frames of presentation. We also find that mass media is politically biased, i.e., it covers certain political parties and their candidates more than others. While [43] find that these biases are dependent more on the criticisms that vary among parties covered, we do not perform analysis of the aggregate stance with which the parties are covered. Our analysis simply suggests that some political parties are over/under-covered by certain mass media houses. We intend to also observe the stance or sentiment alignments with which the parties are reported by news-sources in future.

We also see that the aspect coverage of social media is significantly similar to that of mass media. Hence, we argue that social media aids in furthering the biases existent in terms of aspect coverage in mass media. These findings are in line with [174] which also observes amplification of mass media biases by social media. However, this indication is not as strong when we compare the sentiment slants of mass media articles and the Twitter posts that share them. This signifies that although the users prefer to share the aspects covered by their favorite news-sources, they tweet with a slightly neutral sentiment slant as compared to the news-sources. Finally, although we find existence of echo chambers among follower communities of mass media houses similar to [26], we

also see some signs of users trying to diversify their news consumption by following news outlets belonging to opposite affiliations (as seen from overlap between communities of DecH and NIE). We intend to carry this research further by pursuing a more fine-grained approach of ideology detection of mass media outlets, by analyzing their stance (pro/anti policy) instead of sentiment slants of their articles. We also want to replicate our analysis on a larger dataset of social media users, which also includes users that are not followers of any of the mass media outlets that we study.

This study is a combination of automated computational and qualitative analysis. Hence, it demands a significant amount of manual effort, especially in naming the aspects, building the coding schema, and mapping the aspects to the five constituencies based on the coding schema. Although the size of data analyzed manually is manageable, similar to any qualitative approach, our techniques may suffer from subjectivity. We have attempted to address this issue through multiple rounds of due deliberation, and by calculating metrics like inter-tagger agreement. In future, we also intend to perform the manual analysis using crowd-sourced techniques. Another limitation of our social media analysis approach is that we map only those tweets to the mass media aspects that contain mass media article URLs. This leads to missing out on a lot of tweets, which might be on the policies, but do not contain any URL. As part of our future work, we will try to develop a better approach of mapping the tweets to their aspects.

Although we study economic policy issues as presented in mass media and social media using our framework, it is easily generalizable to any other domain of bias analysis, for instance, analysis of bias in policy documents, political speeches and debates, news belonging to other domains apart from politics, etc.

Chapter 6

Analysis of Discourse on Economic Policies

The parliament, the media, and the citizens, are key participants in any democracy. In this chapter, we study their priorities pertaining to four economic policies in India - Demonetization, Aadhaar, GST, and Farmers' Protests - by examining the content of questions asked by politicians in the parliament, news articles published in the mass media, and data from social media (Twitter). The broad research question that we try to answer through this section is: *Is the policy-making process democratic, i.e., one ensuring equitable representation of all sections of people and their problems?* The method of extracting the aspects for these policy events (and their naming) has already been described in chapter 2. The number of aspects for each policy event is the same as stated in the previous chapter. We now describe the relevant works in this area and our approach to answer the aforementioned question.

6.1 Related Work

To understand if the policy-making process is democratic, it is important to see if aspects related to all sections of people are provided equitable and unbiased representation by the participants of democracy, especially the mass media and the Parliament. We have already discussed about works that show how the mass media and social media carry biases in terms of aspect coverage. Bias in the coverage of important issues in the Parliament has also been analyzed by several researchers, in terms of the questions asked on policies by parliamentarians (MPs).

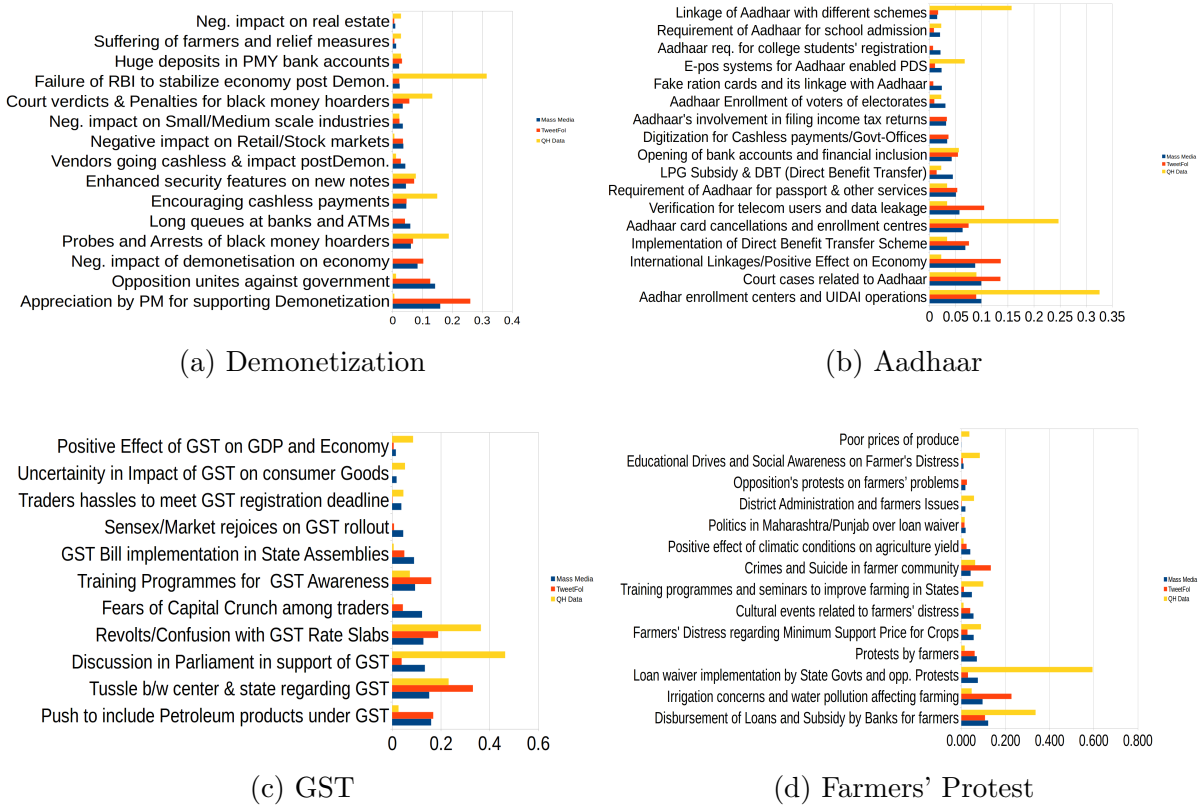
These biases can be inherent and depend on the class or section to which the MP belongs. For instance, Bird [37] studies the effect of asking questions to address gender related concerns by the members of the Parliament, and uses quantitative and qualitative studies to find that MPs generally ask most questions about their own gender. Sometimes, these biases depend on the demography of the constituencies of the MP. For example, Saalfeld [173] studies a set of over 16,000 parliamentary questions tabled by 50 British backbench Members of Parliament (MPs) in the 2005–10 Parliament, and finds that all British MPs respond to electoral incentives arising from the sociodemographic composition of their constituencies – minority and non-minority MPs ask more questions relating to minority concerns, if they represent constituencies with a high share of non-White residents. Biases can also arise in the Parliament based on the importance of the political party to which the MP belongs, and the political/electoral needs of the MPs. On these lines, Ayyangar et al. [20] study Question Hour data of the Indian Parliament for over 30 years, and find that despite increasing importance of sub-national parties in electoral and executive arenas, national parties dominate legislative oversight, and that there is a huge gap in the number of questions asked by these two types of parties. Blidook and Kerby [38] study data compiled from the 34th–37th Canadian federal parliaments, and build a negative binomial regression model of parliamentary question-asking, which demonstrates that Canadian Members of Parliament are both socialized into their roles and that they adapt their behavior in the Parliament to meet their electoral needs.

Our work on the parliamentary questions in India vary from the aforementioned papers in some ways. While these papers focus on studying the parliamentary questions independently, we study parliamentary Question Hour data in conjunction with mass media data. We map the parliamentary questions on a policy to the mass media aspects belonging to it, and see how these aspects are covered in the Parliament.

From the aforementioned studies, we see that there are biases in the coverage of aspects that the parliamentarians cover across geographies. These studies motivate us to ask the research question *Which aspects about the policies do the mass media, the social media community, and the Question Hour data cover?* in the Indian context. The answer to this research question will aid us in understanding if the policy process and especially the questions asked on policy issues represent issues of all sections of people equitably. We also want to see if these biases are segregated on political lines, i.e., if certain parties cover a few issues more than others. Hence, the second research question that we try to answer in this work is *How do the statements or questions vary across the dominant political parties?* In the subsequent sections, we elaborate on these research questions.

6.2 Aspects Covered by the Media and the Parliament

We try to answer the research question: *Which aspects about the policies do the mass media, the social media community, and the Question Hour data cover?* in this section. We analyze each policy separately, to understand the relative aspect coverage given by mass media, social media, and parliamentary question hour data to different aspects. The notion of relative coverage helps us study the *agenda setting* effect of mass media, with respect to the attention it provides to the various aspects of a policy. In figure 6.1, we show the distribution of aspect coverage for mass media, social media, and QH data. Here, we present our analysis on each policy, to see which aspects get what coverage in mass media and QH data, and the implications of the trends observed. The detailed analysis



(a) Demonetization (b) Aadhaar

(c) GST (d) Farmers' Protest

Figure 6.1: Relative aspect coverage of each policy by mass media, social media community, and QH data

can be found in our recently published paper [181]. The aspect coverage distributions for all policies are significantly different from the zero vector ($p - value < 0.005$), which indicates that the mass media, the social media, and the Parliament involve in significant discussions and reporting on various policy aspects. The t-test between the aspect coverage distributions of the three participants of democracy reveals that the distributions are statistically insignificant ($p - value > 0.05$), indicative of the proximity in average behavior of the participants, in terms of policy discussions.

Demonetization We find that the mass media provides high coverage to the govern-

ment's narrative justifying Demonetization to the people, and the arguments about this justification between the government and the opposition. On the other hand, the QH data talks more about the mechanisms to realize the intended goals of the government (like curbing graft money and corruption), and to restore normalcy post implementation of the policy move. The highest covered aspect in the mass media talks about the prime minister's statements on the policy, and his encouragement to the general public and the poor to support the policy move to fight black money and corruption. For QH data, we find that the highest covered aspect is quite different from that of mass media. QH data shows the highest coverage to [*Failure of RBI to stabilize economy post Demonetization*], which includes questions on issues like the menace of fake currency that circulated immediately after the move was announced and overall issues related to currency management.

We thus see that the Parliament acts as a forum to provide feedback on the policy, and does ask relevant questions on improving the situation post its implementation. However, since Demonetization was a highly politicized issue, the mass media covers mostly the politics involved, and covers the state's narrative in support of the policy and the arguments arising from the opposition thereof. Thus, the reference to the poor and the middle class in mass media is only incidental, because the politicization of the issue included narratives and counter narratives built around these two constituencies.

Aadhaar: From our analysis, we find that in Aadhaar, the mass media covers legal issues such as privacy and legitimacy of linking Aadhaar with welfare schemes, and practical issues with implementation, such as problems with enrollment centers. The QH data talks about the same issues as well. However, in both cases, there is more focus on issues of the middle class than of the poor. For example, the enrollment centers talked about are mostly cater to the middle class people in cities. The middle class constituency even gets more attention in the legal discussion topics like privacy, than immediate issues to do with welfare scheme linkages. In fact, as reported in some sources [140], privacy concerns are not as important for the poor as a significant section of poor people like Aadhaar because it gives them an identity. Their problems primarily concern the implementation issues with the policy. The highest covered aspect in Aadhaar by mass media talks about the

court cases related to the policy – primarily on data security, right to privacy, and linking of Aadhaar to welfare schemes. For QH data, this is also the highest covered aspect, which covered questions dominantly around the UIDAI operations and implementation issues related to the Aadhaar scheme, especially on the inability of people to register for Aadhaar due to issues with biometrics during the different phases of Aadhaar implementation, and steps taken to deal with these problems.

We thus find that the QH data again includes technically detailed questions on the policy operations, and on improving its implementation. However, the focus is more on the middle class issues and not the poor. In mass media too the attention is more on informing the middle class citizens about the facilities for connecting them to the policy, the operations around policy implementation, and loopholes related to privacy of data. There is a conspicuous lack of attention to the issues impacting the poor directly.

GST: We find that both the mass media and QH focus on operational issues of GST relevant to the small traders, and not sufficiently on how the common people will get impacted. Therefore, while in Demonetization a narrative was built on how it would benefit the poor and middle class, for GST there was no such attempt made by the politicians, and the mass media also did not see much coverage about such aspects. The top aspect covered in mass media includes discussions on the protest by state politicians supporting inclusion of petroleum products under GST. The rationale behind this demand was that the price of petroleum products would reduce, if they come under the GST umbrella. This is an aspect relevant to the consumers directly.

Other examples of issues related to consumers include rise in price of commodities and services because of more of them coming under the tax umbrella due to a push towards formalization, leading to informal enterprises dying out and making way for formal enterprises. However, these issues do not see much coverage from mass media. The second highest covered aspect in mass media discusses the implementation of the GST bill in different states, with significant amount of discussion revolving around the applicability of GST in Jammu and Kashmir, a state granted special status. This aspect does not relate to the issues of the middle class and the poor directly. For QH data, we find that

the top aspect is again different from the aspect covered most by mass media –[*Discussion in the Parliament in support of GST*], which includes questions regarding administrative issues regarding smooth passage of the GST bill. This aspect does not directly relate to the middle class and the poor, and involves technical discussions on the bill. The second highest covered aspect, [*Objections and confusions regarding GST rate slabs*] includes questions about applicability of GST to various sectors, and the parliamentarians asked most questions relevant to the small traders. Most questions relate to the different rates of GST applied to different sectors of trade. This aspect does not directly connect to the middle class or the poor as well.

Therefore, as stated earlier, in GST the focus of the Parliament was primarily on the traders and companies, on push towards formalization, and not on how the poor and middle class consumers will be impacted because of GST. This trend provides an indication that GST was seen mostly by the government as a source of tax revenue through formalization, without much consideration given to the impact on the consumers [187]. This trend is also carried forward by mass media where the troubles faced by traders in registering for GST and political discussions on GST come among some of the top aspects covered. However, the mass media also gives highest coverage to the issue of bringing petroleum products under GST, which is directly relevant to the consumers. We also find that neither the mass media, nor the Parliament provides attention to the informal sector.

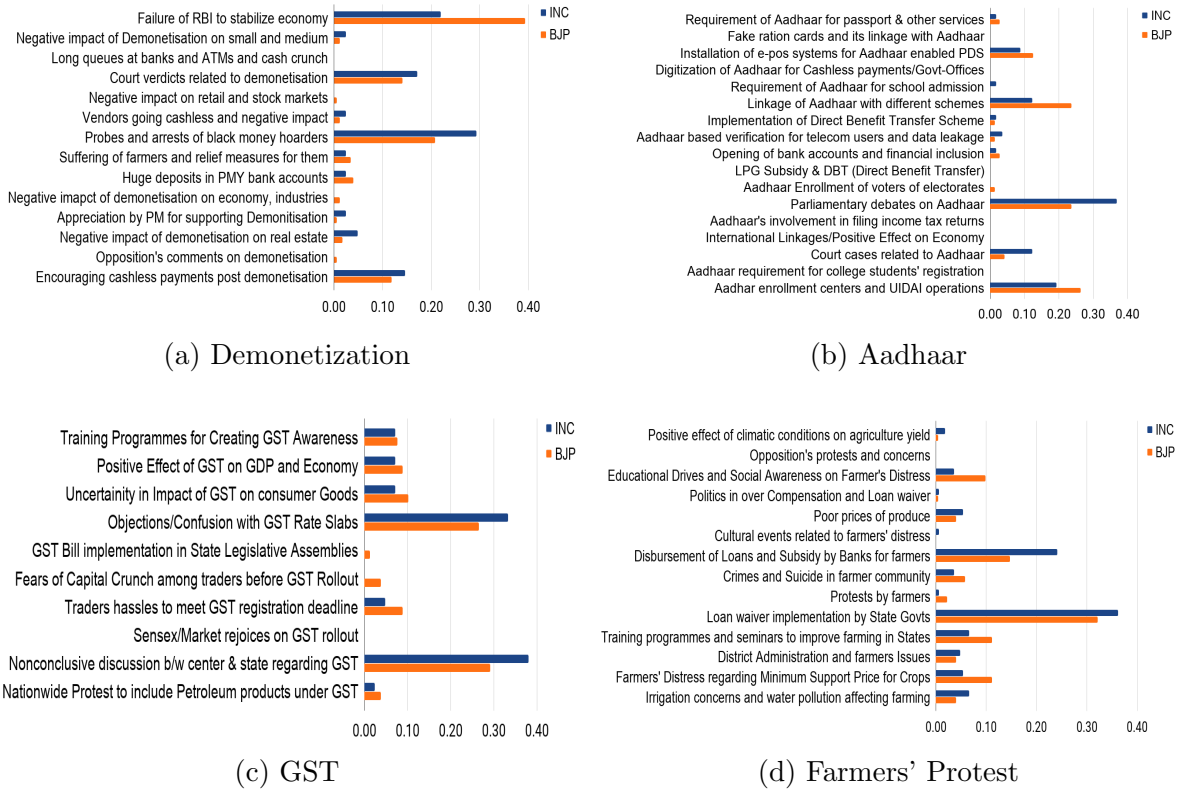
Farmers' Protest We find that in Farmers' Protest, both mass media and QH focus on loan waivers, with the mass media discussing about the reasons behind the poor farmers' debts, and structural changes required in the agricultural sector [163], while the QH discussing the mechanics of loan waivers and implementation of loan programmes by states. [*Disbursement of loans and subsidies to farmers by banks*] is the highest covered aspect in mass media for Farmers' Protest, and talks about the loans provided to the farmers by banks, under various government schemes. In mass media, the second highest covered aspect is [*Irrigation concerns and water pollution affecting farming*], which talks about the problems faced by farmers in irrigation, due to water pollution through industrial

effluents. The QH data shows highest coverage for the aspects [*Loan waiver implementation by state governments and oppositions' protests*] and [*Disbursement of loans and subsidies to farmers by banks*]. The questions asked in both of these aspects concern the technicalities around the implementation of different loan waiver schemes and the problems around them. However, unlike mass media, they do not sufficiently address the structural issues related to the agricultural scenario, or ask about the root causes behind the farmers' distress (like irrigation concerns and water pollution and issues with minimum support price of crops), or their solutions (like generating awareness on farmers' distress, and training programmes to improve farming through advanced and scientific methods).

Hence, although the mass media does cover the structural issues in agriculture, this trend is not carried forward in parliamentary discussions. The QH data shows the tendency of policymakers to achieve a quick remedy to problems of the farmers by supplying them loans through various loan schemes. However, we do not see sufficient attention given to the structural issues causing problems for the farmers or the appropriate solutions to these issues. Thus, there exists a lack of detailed understanding of their issues and the structural changes required to address them in the Parliament.

6.3 Variation in Questions Asked by Political Parties

Here we see how the coverage of aspects corresponding to the four policies varies across the two biggest political parties in the QH data by answering the research question: *How do the statements or questions vary across the dominant political parties?* We present the coverage given to each aspect for each policy by the BJP (406 MPs considering LS'15 and 16) and INC (292 MPs considering LS'15 and 16) in figure 6.2. We report our detailed results in [181]. We see that despite having a much smaller number of MPs compared to BJP considering both LokSabha terms, INC asks a significant number of questions on each policy. The t-test between the aspect coverage distributions of BJP and INC reveals



(a) Demonetization

(b) Aadhaar

(c) GST

(d) Farmers' Protest

Figure 6.2: Relative coverage of aspects provided by political parties in QH data for the four policies

that the difference between their means is statistically insignificant ($p - value > 0.5$). This indicates that both BJP and INC indulge in similar extent of questioning in the Parliament. INC is seen to be questioning the policies, especially regarding the procedural aspects and mechanics around their implementation. It is also evident that except for Demonetization, aspects related to the poor and the middle class, some of which address their issues in-depth, do see a coverage from the political parties. However, their overall coverage is much smaller when compared to some of the highest covered aspects, which do not analyze the immediate problems of the poor in-depth or address them at all (for Aadhaar and Farmers' Protests), or the problems of the consumers (for GST).

6.4 Discussion and Conclusion

Our work attempts to answer the question of how representative in its deliberations the Indian policy-making process is. In other words, we try to find out if the policy discourse in mass media, in social media, and in the Parliament provides equitable attention to the concerns of all sections of citizens, i.e., if the democracy is representative. Our work also hints towards the deliberativeness of democracy, i.e., what is the depth of discussions that ensue on issues relevant to the different sections that the participants represent.

The research question that we tried to answer in this section was *Is the policy-making process democratic, i.e., it ensures equitable representation of all sections of people and their problems?* We find that the mass media covers stories about different constituencies (the poor, middle-class, corporations, etc.) but shows biases in terms of which constituencies it chooses to focus on for each policy. The parliament tends to focus on procedural aspects of the policies. We also find that the social media simply echoes the trends of whatever is emphasized more in the mass media, without much deviation in the attention placed by social media users on different issues. We further find out that instead of selecting issues of importance based on feedback from citizens, the parliamentarians mostly indulge in partisanship, and shape their questions based on party goals that change depending on whether their party is in power or not. Overall, we are able to use this analysis to state that there still lies enough scope for the participants of democracy to make the policy-making process democratic.

Overall, our findings suggest that the participants of Indian democracy are not sufficiently representative. There are policy events for which entire constituencies are ignored by all of the participants of democracy – for example, in Aadhaar, we find a conspicuous lack of attention provided to the immediate issues of the poor by both the mass media and the Parliament. Even when the constituencies are covered in the policy discourse, many a times it happens by way of politicization (especially in the mass media) where politicians indulge in blame-games with each other around that constituency, instead of focusing on constructively addressing the issues or on providing a nuanced viewpoint. Several

reports [120, 111] on the role of news media in India focuses on how the Indian mass media fails to cover nuances of genuine issues important to the public or to the poor. This lack of representativeness is also exhibited by the Indian Parliament as argued by existing studies. Kapur et al. [114] state that the profusion of political parties in the Indian Parliament has substantially increased the barriers to collective action, owing to increased politicization of the issues debated. Wallack [221] argues that the Indian Parliament is not acting as a sufficiently representative institution for the people, and focuses on the frequent disruptions and lack of general interest regarding the issues of concern.

Our findings also indicate that the participants of Indian democracy have a scope of improvement towards being deliberative. We find that there are policies like Farmers' Protests for which the structural issues that actually should be discussed in the mass media and in the Parliament are nearly neglected. Instead, surface level criticisms and quick fixes appear most often on these policies. Glaser discusses in his book *Experts Versus Laymen* [86] that experts (in this case policymakers) often have a wide gap from laymen (citizens) in terms of the power and expertise that they carry, and thus, the laymen are mostly unable to shift the balance of power towards themselves, and make the experts accountable towards the decisions that they make. This effect is clearly seen in the Indian democracy – the citizens/beneficiaries are often unable to provide feedback to the policymakers and to participate in the policy discourse. This results in lack of accountability on part of the policymakers, hence leading to a deficiency of attention to the structural issues of concern.

Chapter 7

Analysis of Policy Representation in Mass Media

policy-making is influenced by a number of factors, including electoral politics, ideological biases of actors involved in the policy-making process, bias in mass media, and the interlocks between corporate and government entities. In this chapter, we study four technology or ICTD policies in India, and explore the political economy around them by using data about how these policies are covered in the mass media. We study which actors are covered more in media, how they speak on the policy issues, and which aspects are given more coverage for these policies. This analysis helps us answer the broad research question: *How are policies justified through the mass media in India, and by whom?* In the subsequent sections, we elaborate on the sub-research questions we ask in this regard.

The ICTD policies that we study are Aadhaar [226], Digital India [224], Cashless Economy [46], and National E-Governance Plan [225]. Articles are initially collected by the media module based on a set of manually selected seed set of keywords related to the event as shown in table 7.1. The article extraction technique using the keyword augmentation process is already described in chapter 2. We next run LDA on these articles. The accuracy of LDA mapping for the four ICTD policies lie in the range of 74-78%. In the

Keywords (manually selected)
Aadhaar: aadhar, aadhaar, adhar, adharcard, aadharcard, aadhaarcard, uidai, aadhar card, public distribution system, pds, ration card, ration, e-pos
Digital India: digital india, digital swades, india digital, digit india, digital desh, make in india, digital divide, digital payment,free wifi service, digital locker, digital transaction, wifi hotspot, budget cybersecurity, skill india, internet connectivity, smart city, digital business, bharatnet project, digital present, bharatnet, digitalised, digitalized
NeGP: e-governance, information and communication technology, e-govt, e-government, electronic governance, paperless office, communication technology, ict academy, ict sector, ict information, ict tool, e-district, m-governance
Cashless Economy: cashless, digital payment, mobikwik, unified payment interface, upi, online transfer, sbi pay, icici pocket, payzapp, freecharge, e-wallet, mobile wallet, internet banking, net banking, mobile banking, PhonePe, physical-POS, M-POS, V-POS, digital transaction, pos machine, swipe machine, digital wallet, digital economy, card payment, bhim, bhim app, banking transaction, swiping machine,payment gateway

Table 7.1: List of manually collected keywords used to extract articles (and tweets) corresponding to the ICTD policy events. Here, we only show the manually selected keywords after converting them to lowercase, and after pre-processing of the articles was done.

subsequent sections, we answer some of the research questions that we ask in this study.

7.1 Related Work

In this section, we describe some of the related studies that motivate our analysis of policy representation in mass media, with respect to four policies related to Information and Communication Technologies for Development (ICTD). It is evident that in the current times, there is a significant push towards development and adoption of ICTs globally, and the trend is the same in India. As reported by prominent media houses, the ICT spending in India is on a constant rise [45], and there is a consistent push from the state towards use of these ICTs. On the other hand, big business houses and influential business-persons are also seen to promote these ICTs and their advantages. We present some existent studies

that focus on the political and mass media discourse on technologies, and corporatization of mass media.

Corporate-Political discourse on technology: With every new technological intervention in the policy space, we see an increasing propensity of the policymakers to project advanced technology as the solution to all problems. This trend is supported by several theories. James C. Scott in his book, *Seeing Like a State* [177] calls this tendency of the state to favor scientific or technological advancements in every field of human activity *high modernism*. The author also provides arguments on how the state exhibits use of force to implement these designs that it feels is indispensable for societal development. David Harvey [95] calls this overemphasis placed on technology to bring social change *The Fetish of Technology*. He argues that capitalist entrepreneurs and corporations focus on innovating newer technologies everyday not only because they want to do so, but because they have to in order to either acquire or retain their status as capitalists.

These theories encourage us to understand if this trend of high-modernism is prevalent in technological policies in India as well. While we majorly study economic policies in this thesis where our focus is on analyzing the representativeness of the policy process and the biases in various participants of democracy around these policies, we also want to observe the political economy around technological (or ICTD) policies. To be more precise, we want to look at the corporate-political narrative around these policies and the way these technological policies are promoted by corporate-political entities, given that their high-modernistic viewpoints have already been studied in various works.

For instance, the implementation of *Aadhaar* by the Government of India has been criticized due to this approach of high modernism by economists and policy experts in India [68, 205]. In these criticisms, the inability of the technical intervention behind Aadhaar to solve the basic problems of common people has been stressed upon. However, despite these criticisms, Aadhaar survived and the enrollments under the scheme are adding up everyday. Janaki Srinivasan [197] attributes this survival of the Unique Identification (UID) project to three factors – the team, the artifacts, and the benefits projected. She states that *Nandan Nilekani*, the then IT-entrepreneur chairperson of UIDAI proved to be

a popular image among the adult, educated masses (the team) and resulted in promotion of Aadhaar. The concept of getting a 'state acknowledged ID' and its technical prowess (the artifact) – like linking it to financial services – also appealed to the masses. Finally, the benefits of the scheme like elimination of poverty and corruption, as projected by the politicians resulted in convincing the public about the scheme's usefulness.

Pal et al. [153] similarly conduct a study of 200 shopkeepers in Mumbai and Bengaluru, to see how adoption of cashless technologies evolved post Demonetization. They find that in order to justify the policy move, the Prime Minister increasingly emphasized on the usage of digital cash and payment wallets by invoking patriotism, technical advancement, and projecting cashless payment as a one-shot solution to the problems of people. They also find that although immediately post Demonetization, the use of cashless mode of payments rose in businesses, it was followed by an immediate drop after sufficient cash was available again. The authors argue that although cashless payment was force-fed to people by the state in the wake of Demonetization, there were several barriers to quick adoption of technical interfaces by people like technophobia, misinformation, interface challenges, and inertia towards old payment practices. Pal et al. [152] similarly discuss how the current Prime Minister of India branded his image as a tech-savvy modernizer on social media and other platforms, and justified policies like Demonetization by building a narrative around cashless economy and technological advances.

Several researchers have pointed out the problems with this high-modernistic discourse. Pal et al. [152] and Heeks [98] show how there is always ever increasing gap between its supply-side acceptance and demand-side acceptance of technological interventions. There are also problems of corruption, which lead to failure of these interventions at multiple levels [30]. Pal [151] also emphasizes that technology is only a very small part of the challenge of social inclusion, and how the public sphere should evolve in order to be more inclusive of all sections of people.

The aforementioned papers in this section motivate us to understand what and how the policymakers and other corporate-government entities speak about a technological policy issue, since their statements have a great influence on public opinion regarding the policy.

Additionally, we want to understand if there exists a disparity between mass media's coverage provided to different types of entities (selection bias). For example, a very high coverage provided to ruling party politicians as compared to the opposition or civil society members might influence the policy discourse significantly, since it might lead to a one-sided viewpoint being conveyed by the mass media. Specifically, we attempt to answer two research questions. The question *Which entities and groups of entities are the most vocal in mass media on policy issues?* enables us to understand which entities speak the most in mass media (or might be covered the most by mass media) on policy issues and the statements they make on them. The second question *What is the sentiment slant of these elites regarding the policies?* aids us in seeing how these entities speak about the policies, i.e., if they support or oppose the policy in consideration.

Corporatization of mass media and Mass media's discourse on technology: As discussed in the last section, several studies have shown that the policymakers and corporate business-persons shape their statements generally in favor of technology. These influential elites may have connections with mass media houses as well, sometimes in the form of ownership networks. Media's dependence on its corporate or political owners may lead to a significant influence on its discourse on technology. This is also argued by Herman and Chomsky [99], who identify five of the filters that decide the content that media produces. Among these five, the three most important filters according to the authors are ownership of the medium (corporate ownership), the medium's funding sources (its advertisers), and its sourcing (the information source accessible to the media).

There exists ample literature on the study of ownership networks (especially corporate ownership) of mass media houses. Paranjoy Guha Thakurta in his book [212] states that there are more than 82,000 publications registered with the Registrar of Newspapers as on 31 March 2011, but the mass media in India is dominated by less than around a hundred large groups or business conglomerates. On similar lines, Arsenault et al. [19] study multi-media business networks across the US and see how the major media corporations are interlocked with highly influential business groups and politicians. Specifically, they show that these media corporations are interlocked with networks of finance, production,

advertising, technology, research, and politics through multiple switches. In a separate work Arsenault [18], studies the operational dynamics of Rupert Murdoch and NewsCorp to show how corporate media entities negotiate the power dynamics of the network society to support its business goals. It focuses on the chief strategies to do so, including political brokering, leveraging public opinion, formulation of sensationalist news formulas, customizing media content and diversifying and adapting media holdings.

Apart from the studies of ownership network structures of mass media, there also exist studies that provide conclusive evidence of corporate influence on the content presented in mass media. Gilens and Hertzman [85] show that the Telecommunications Bill, formulated to loosen the cap on Television channel ownership in the US, received very different responses from popular mass media houses, depending on whether the policy benefited the mass media owners or not – news-sources that could gain from the policy provided favorable coverage to it, while the coverage of this issue in newspapers owned by companies that did not benefit from the policy was unfavorable.

Another way to study the corporatization of mass media is to analyze media advertising trends. In this domain, Focke et al. [77] prove that in the US scenario, advertising does bias the content presented by the news-sources – the news-sources especially report less negatively about the advertisers, in case of an adverse event or scenario involving the advertisers. On similar lines, Smet and Vanormelingen [62] argue that newspapers in Belgium provide higher coverage to firms that advertise in them and also provide these firms a positively biased coverage. They, empirically, deduce the exact amount a firm needs to spend per month on the newspaper, in order to get one extra mention in the same month. Although in our work we do not link the content presented or entities covered in the policy discourse with ownership networks of media houses, the aforementioned papers motivate us to study how the Indian mass media covers the discourse on technological policies. An alignment between the aspects highly covered in mass media and the ones favored by the corporate-political entities can help us understand whether the Indian media is becoming a mouthpiece for these elites regarding technological interventions, or if it is an independent and unbiased platform for policy representation.

Mass media's discourse on technology has been studied by several researchers. Cacciatore et al. [47] study the thematic coverage related to two technology topics, namely nanotechnology and nuclear technology, in both online news-sources and print news between 2004 and 2009. They observe the coverage of various themes and sub-themes on these technologies, and conclude that online news aggregators (like Google News) provide a much varied and balanced coverage to technical topics than print news. On similar lines, Sengers et al. [183] study the media narrative around 'Biofuels' between 2000 and 2008 in Netherlands. They use a combination of quantitative content analysis of the text in mass media, and discourse analysis of the statements made by Biofuel practitioners. The authors observe how the narrative has shifted from positive impact of the technology to its negative sides (like environmental concerns) over time, through the themes captured corresponding to the narrative. Mallett et al. [123] study the Canadian mass media's discourse on *smart-grids* between 1990 and 2012. Similar to our work, the authors in this paper perform a qualitative coding of different types of articles on the topic to identify their frames. Studying the evolution of these frames, they find that the discourse has shifted towards the negative side over time, and also that the risks and benefits of the technology are location specific.

Another work that motivated our research in understanding the political economy around technological policies is that of Nisbet and Lewenstein [142]. In this work, the authors study the media coverage of Biotechnology articles between 1970 and 1999 in the US, and study the coverage of themes, the frames of presentation, the tone with which the articles are written, and the chief actors covered frequently by the media. They find that the coverage of Biotechnology over time is primarily positive and high-modernistic, with minimal discussion of the risks involved. Unlike the other studies, this paper also analyses the entities covered in the policy discourse, and points towards a hegemonic coverage of policymakers, scientists, and industrialists by the mass media. While the research questions asked in this paper align with our work, their approach is mainly qualitative where the authors code media articles on dominant themes and frames.

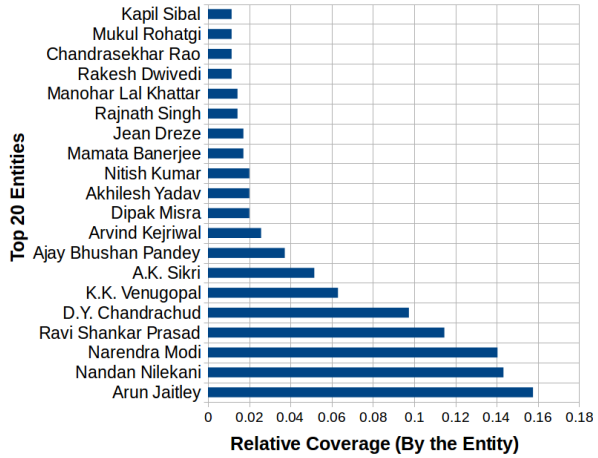
Cacciatore, Mallett, and Sengers [47, 123, 183] also discover themes by a keyword based

approach. While we too analyze the themes or dominant aspects in the media discourse, our automated discovery of aspects using LDA is different from their keyword based or qualitative approach of discovering themes – while our method aids in discovering any latent aspect in a policy issue, a keyword based approach will lead to discovery of content related to a static set of themes. Additionally, unlike the aforementioned papers, we do not study a specific technological policy but cover a range of ICTD policies, which makes our domain of study much broader.

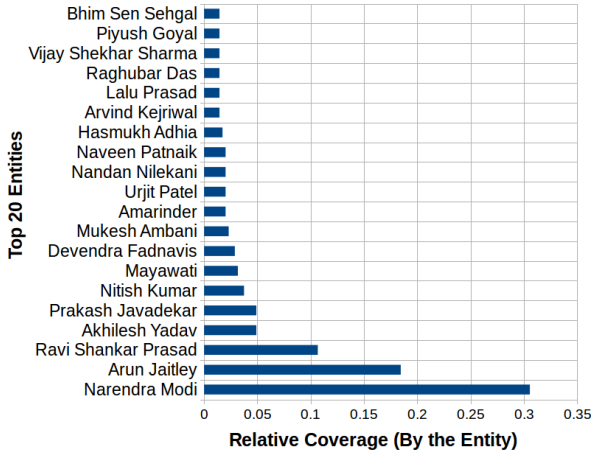
The studies mentioned in this section motivate us to answer the research question: *Which aspects are covered more by mass media on ICTD policy issues?*. While we discuss the political discourse on ICTD or technological policies in this work, this question enables us to see how the mass media in India, as an avenue of information, project technological policies to the public, i.e., if it's aggregate coverage on technology aligns with that of the government and corporate elites.

7.2 Most Vocal Entities and Groups in Mass Media

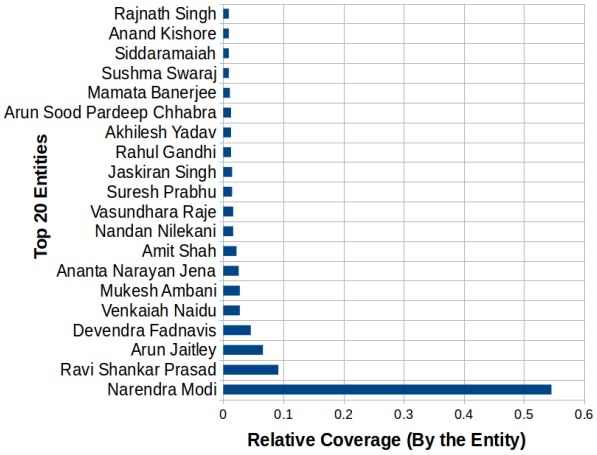
In this section, we attempt to answer the research question: *Which entities and groups of entities are the most vocal in mass media on policy issues?* To answer this question, we calculated the relative coverage of each entity in our mass media dataset, and ranked them in descending order of their relative coverage. This gives us a ranked list of entities that are mentioned the most in media, corresponding to the policy issues. Figure 7.1 shows the plot of relative coverage (for statements made by the entities) for the top 20 entities with highest relative coverage. The plots reveal some interesting trends. We find that most of the entities with top relative coverage in all of the policies considered are politicians. The t-test between coverage distributions of the top 20 politicians and other type of entities (directors, judiciary members, and bureaucrats) shows that the difference between them is significant for Aadhaar and E-governance ($p - value < 0.05$). For Demonetization and Digital India, the difference between the coverage distributions



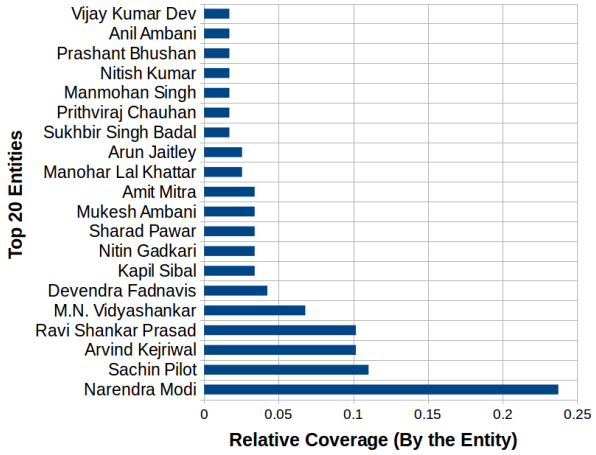
(a) Aadhaar



(b) Cashless Payments



(c) Digital India



(d) E-governance

Figure 7.1: Plot of the relative coverage of top 20 entities for each policy for statements made by them: relative coverage is calculated as the number of statements made by the entity divided by the total number of statements by all entities, corresponding to a policy.

of politicians and directors comes out as statistically insignificant ($p-value > 0.08$). This indicates that most of the discussions around policies see high coverage of politicians and political statements, and as we explain later, this coverage is mostly about political blame-games than an introspection into the technical nuances of the policies. Additionally, directors, bureaucrats, and judiciary members also get a high coverage, next to politicians, in policy matters. *Aadhaar* shows coverage of non-politicians like *D.Y. Chandrachud, K.K. Venugopal, A.K. Sikri, and Rakesh Dwivedi*, who are all judiciary members of the Supreme Court of India, which can be explained by the fact that a lot of debates took place in the judiciary around the *Aadhaar* policy, although it revolved around the constitutional legitimacy of the policy. We also find the presence of *Ajay Bhushan Pandey*, a bureaucrat in *Aadhaar* and currently the CEO of UIDAI. The policy push towards *Cashless Payments* shows the presence of business-persons like *Mukesh Ambani and Vijay Shekhar Sharma*[1], judiciary members like *Bhim Sen Sehgal*, and economic advisors like *Urjit Patel and Nandan Nilekani* because it is an economic policy issue. *Digital India* policy too contains *Mukesh Ambani and Nandan Nilekani* among the top covered entities. This is because *Ambani* brought out the Jio network, which has disrupted the telecom space by providing very low-cost 4G connectivity. For *eGov* policy priority, we have business-persons like *Mukesh Ambani, Anil Ambani, and M.N. Vidyashankar* among the top covered entities, who were supporters of the move. Our findings are reported in the paper [180].

Two trends are evident from the data: (a) After politicians, business-persons are provided maximum coverage by the mass media (except in *Aadhaar* where judiciary members get a high coverage), and (b) Academicians and social development experts from the civil society (including social activists and researchers documenting successes and failures of these policies) are provided negligible coverage by mass media (t-test between distributions of politicians and experts shows a $p - value < 0.005$). It should be noted though that the articles analyzed here are news articles, which do not include opinions and editorials (opeds). On considering only opeds, we find a higher coverage provided to academicians and policy experts (around 8%, 8%, 0.6% and 13% for *Aadhaar, Cashless Economy, Dig-*

[1]Vijay Shekhar Sharma is the founder of PayTM, which gained immediate leverage in the wake of Cashless Economy.

ital India and E-governance, respectively.). However, this coverage is still much smaller compared to political parties (57%, 51%, 85%, and 76% to BJP for *Aadhaar, Cashless Economy, Digital India, and E-governance*, respectively.).

We find consistent coverage of influential business-persons like *Mukesh Ambani and Nandan Nilekani* across all of the four policies. On analyzing the statements made by these business-persons, we find that they mostly talk about the positive aspects of technology led change with respect to these policies. For example, *Nandan Nilekani's* statement [65] on *Aadhaar*: *"You need to keep people's healthcare record electronically because health records could be voluminous because you have x-rays, MRIs and ultrasounds to be stored and you have to do digitally"* and *Mukesh Ambani's* statement [66] on Cashless Economy: *"... the company is planning to launch a digital marketplace for its b2b business, including the kirana shops, across the country."* are indicative of an ideology of technology driven change. Similarly, top politicians like *Narendra Modi's* statements [217] on Cashless Economy too reflect a technology driven developmental ideology: *"When poor farmers of villages have started adopting digital payment, now they (middlemen) have started spreading new rumours."*. On the other hand, views such as those on the failure of *Aadhaar* implementation leading to denial of ration to poor, by development economist *Jean Dreze*, highlighting problems with the policy implementation do not get as much coverage. We find an insignificant presence of academicians and social activists in the discourse on policies, compared to politicians and business-persons. In table 7.2, we show the total media coverage for groups of entities, grouped by their political parties or professional backgrounds. At the party level, we see that, as expected the ruling party, BJP, is given the maximum coverage by the media across most policies. INC, the main opposition, comes next in terms of coverage (and has a much lesser coverage percentage than BJP). This is followed by bureaucrats, business-persons (directors), and celebrities. We again see that negligible coverage is provided by mass media to academicians (especially economists and policy experts) across all of the four policy issues. Among the non-political entities, the high coverage given to bureaucrats and business-persons can be justified as these economic policy issues require an active involvement by them for implementation, but similarly the role of civil society which is a crucial pillar too in

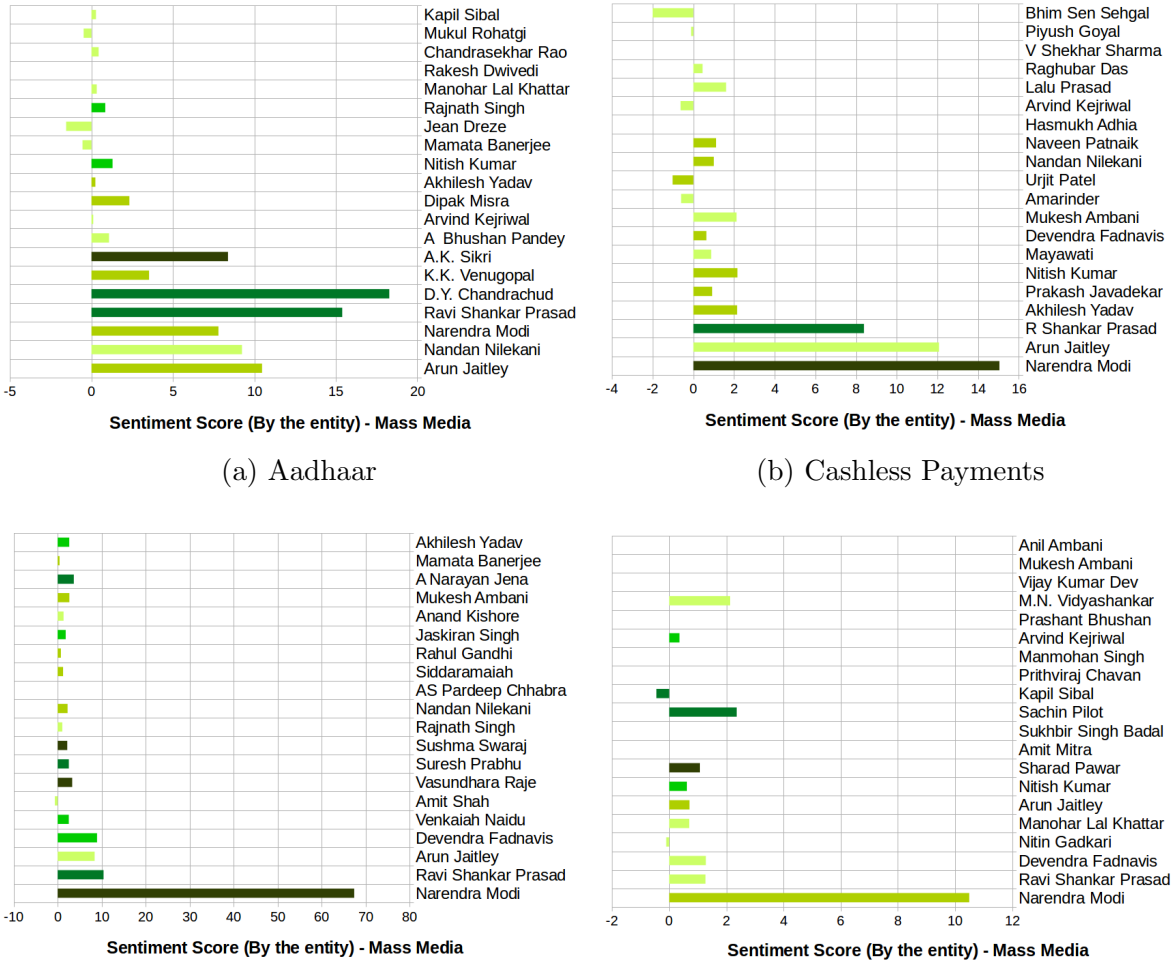
	Aadhaar	Cashless Economy	Digital India	E-governance
BJP	24.81	**57.30**	**78.61**	**57.31**
INC	**25.18**	14.62	5.91	14.62
Bureaucrats	2.51	7.21	2.59	7.21
Business-persons	0.66	2.73	2.69	2.73
Celebrities	1.40	1.17	1.87	1.17
Academicians	7.09	0.00	0.06	0.00

Table 7.2: Relative coverage in percentage for entity groups (considering both *about* and *by* statements): BJP and INC are the two biggest parties in India (BJP being the ruling party currently).

implementing these policies is hardly mentioned. On similar lines, the relative coverage of celebrities (like *Ramdev*, the yoga guru, and owner of the ayurvedic firm *Patanjali* in relation to the *Aadhaar* policy, and the singer *Mehmood Akhtar* in relation to *Cashless economy*) is more than academicians.

7.3 Sentiment Slant of Statements by Elites

We answer the research question: *What is the sentiment slant of these elites regarding the policies?* To see the orientation of the most vocal entities on the policy issues, we measured the overall sentiment slant of entities towards the policies (across all newspapers) as the sum total of the sentiment scores for all sentences that the entity occurs in. We carried out this analysis for the *by* class, i.e., for statements made by the entities, and show our results in figure 7.2. In the figure, the bars on the right hand side of the zero value on x-axis represent positive aggregate sentiment, and those on the left hand side represent negative aggregate sentiment. We color coded these bars on *degpol*. Darker the color of the bar corresponding to the entity, higher is the aggregate polarity of the entity's sentiment.



(a) Aadhaar

(b) Cashless Payments

(c) Digital India

(d) E-governance

Figure 7.2: Plot of the aggregate sentiment, color coded on *degpol* for the top 20 entities with highest coverage: the aggregate sentiment/degpol is calculated as the sum total of the values corresponding to the statements made by an entity. Higher the value of degpol (darker the color of the bar), more is the overall polarity.

For all policies, In terms of aggregate sentiment slant, *Narendra Modi* is consistently seen to have a high positive sentiment score, except in *Aadhaar*. *Arun Jaitley* and *Ravi*

Shankar Prasad are also seen to have a quite high aggregate sentiment slant except in E-governance. We can also see that in *Aadhaar*, the aggregate positive sentiment is well distributed across several entities. On the other hand, it is skewed towards Modi (and his two ministers *Ravi Shankar Prasad* and *Arun Jaitley* in some cases), and nearly insignificant for other entities for the other three policies. This indicates that Modi being the prime minister and the most popular face of the current ruling party, made maximum number of positive comments on these policy issues, and was also covered highly by the media. Compared to his media presence, the coverage given to the other entities is much lesser (especially in Digital India and E-governance). Moreover, *Aadhaar* being the most widely discussed policy among the four policy issues, received comments from a lot of entities belonging to different sectors. This wide participation from various sectors was not reflected as much in the other policies, which mainly received attention from politicians.

We also see that the aggregate sentiment slant for all of the policies under consideration is mostly positive. This is because the actors covered by the media were mostly politicians; the ruling ones having generally positioned these policies as being good for development, and the opposition ones also being supporters, since several of them had been initiated when the opposition was in power earlier. *Aadhaar, Digital India, and E-governance* are policies that were all initiated by the INC when it was in power. *Cashless Economy* was however initiated by the currently ruling party BJP, and saw the opposition having more negative comments, which is an exception to their otherwise mostly positive coverage. Some judiciary members, opposition party politicians, and bureaucrats do have a slightly negative slant, but since the coverage given to them is much lesser than that given to politicians, these views are hardly able to become mainstream. These findings are also supported by the t-test carried out between the aggregate positive and negative sentiment distributions and the zero vector. Except *Cashless Economy*, the negative sentiment vectors for all policies show statistically insignificant differences from the zero vector ($p - value > 0.05$). On the other hand, the difference between the positive sentiment distribution and the zero vector is statistically significant across all policies ($p - value < 0.05$). These findings indicate the absence of significant negative comments,

and a significant positive coverage from the entities, for most technology policies.

In terms of polarization (*degpol*), *Narendra Modi and Ravi Shankar Prasad* are seen to be consistently polar in terms of their statements made across all of the four policy events. For example, *Narendra Modi's* statement for Cashless Economy (against the opposition's criticisms): *"They will keep abusing from whichever platform they get but we have to take our nation to the forefront of the world."* and *Prasad's* comment on Digital India: *"Indian digital economy is going to touch USD one trillion in next five years, giving enormous business scope for the IT industry, with initiatives on good governance and faster delivery driving domestic demand."* are examples of highly polar statements in the political domain. This is expected as *Narendra Modi* being the prime minister of India, was the proponent and staunch supporter of these policies. *Ravi Shankar Prasad* holds the Electronics and Information Technology portfolio, which is the prime functional ministry for all of these policies.

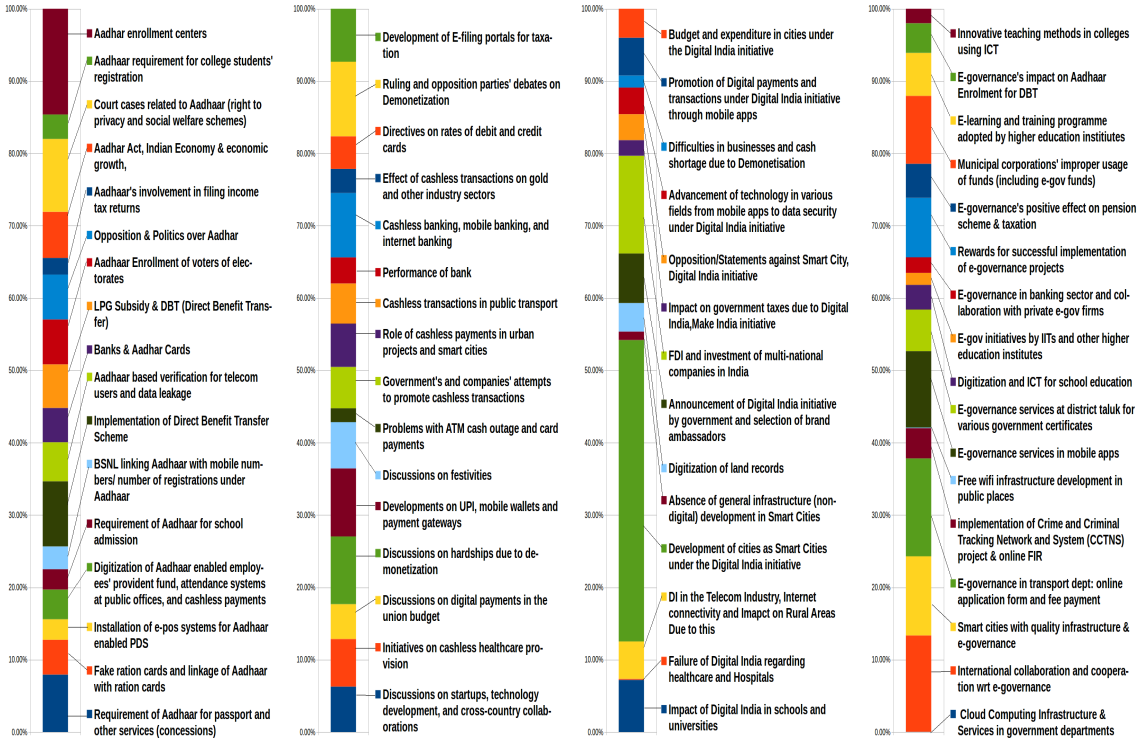
We also find that although the business-persons consistently get much lesser coverage compared to the politicians, they generally speak with a positive sentiment slant towards policies as also explored in the previous research question. For example, in all of the policy events, *Nandan Nilekani* is seen to speak positively. He speaks most positively about *Aadhaar*, which is expected as he was the chairman of UIDAI (the organization that issues Aadhaar numbers to citizens) and the founder of Aadhaar project. *Mukesh Ambani* is seen to speak positively on Digital India and Cashless Payments.

Some examples of aspects on which the most prominent entities in the media spoke, also verify our findings. For example, for *Aadhaar*, we find *Nandan Nilekani*, the architect of the scheme to be less polar than *Narendra Modi*. This is because *Nilekani* mainly spoke on the applications of *Aadhaar* [216] (e.g. *"... its wider application in areas such as passport issuance, online identity verification and attendance in government offices will be seen in the coming days."*). On the other hand, *Modi* primarily engaged in political topics. For example, his statement on *Cashless economy* [217] against the opposition political party, *"They will keep abusing from whichever platform they get but we have to take our nation to the forefront of the world."* indicates his tendency to use nationalism as a factor to

justify the policy implementation. The negative stance of academicians like *Jean Dreze* can be attributed to issues raised by him of starvation leading to deaths, which originated from a denial of food grains in the PDS system from a malfunctioning Aadhaar linkage of the beneficiary family (e.g. *"The state government must clarify about the orders which deny ration on account of Aadhaar seeding and biometrics, and the government should release a white paper to reveal how many people in the state are denied ration due to these reasons.")*[196].

7.4 Top Aspects Covered by Mass Media

Under the research question: *Which aspects are covered more by mass media on these policy issues?* we try to understand if mass media gives selective preference to some aspects (topics) more than others, corresponding to each policy. In figure 7.3, we show the mean aspect coverage for the different aspects corresponding to the four policy events. As we can see from the plots, for *Aadhaar*, the aspect *Aadhaar enrollment centers* has the maximum mean relative coverage, indicative of the mass media's push towards informing citizens about the enrollment centers. The aspect *Court cases related to Aadhaar*, which covers cases on privacy issues of Aadhaar data, and its applicability to public services, has the second largest coverage. These aspects are relevant to the consumer middle class, which is the dominant media audience anyway. For *E-governance*, we find the aspect *E-governance in transport department* to be one of the most widely covered aspects, which covers online applications and technologies built for convenient fee payment by the stakeholders, again an aspect relevant for the middle class. For *Cashless Economy*, the aspect *Ruling and opposition parties' debates on Demonetization* gets the highest mean relative coverage in mass media. This is followed by aspects like *Cashless banking, mobile banking, and Internet banking; Developments on UPI, mobile wallets, and payment gateways; and Discussion on hardships due to Demonetization.* These aspects mostly represent discussions on the advanced technologies and applications implemented for the policy, and the troubles that common people emphasize, but nearly ignores issues faced



(a) Aadhaar (b) Cashless Economy (c) Digital India (d) E-governance

Figure 7.3: Mean relative coverage of aspects corresponding to the four policy events.

by the poor. For *Digital India*, we find *Development of cities as Smart Cities* to have the highest coverage, followed by *FDI and investments of multinational companies in India*. Aspects like *Directives on rates of Debit/Credit cards (Cashless Economy), and Digitization of land records (Digital India)*, which provide people knowledge about the actual technical nuances of the policy issue, are neglected on the other hand.

We find from our analysis that issues related to the concerns of common people, which is the middle class (in *Aadhaar and Cashless Economy*) get significant coverage in mass media. On the other hand, issues related to the poor get much lesser coverage. In general, there is a significant coverage of issues related to futuristic technologies, applications,

technology led change, and role of technical advancement in development (Smart Cities). This trend can increasingly be captured in the form of editorials like, *"The need of the hour is to harness technology and foster innovation for creating a centre of excellence for manufacturing and services in the identified sectors (The Indian Express)."*[158]. Ruling party politicians are also seen to contribute significantly to these editorial spaces and blogs in mass media. For example, Ravi Shankar Prasad's statement on Digital India is clearly seen to be a staunch supporter of technology driven change, *"After coming to power, Prime Minister Narendra Modi gave the vision of Digital India as an important program to transform India through the power of technology and bridge the digital divide (The Hindu)."*[162]. On the other hand, issues covering technical nuances of policies (like *Aadhaar's involvement in filing income tax returns*), issues that analyze failures of policy implementation and technology in general (like *Court cases related to Aadhaar*), and issues of the poor (like *Installation of e-POS and problems with PDS)*) get comparatively less coverage in mass media.

7.5 Discussion and Conclusion

The broad research question that we tried to answer in this chapter was *How are policies justified to the citizens in India, and by whom?* Some of our findings in this section are along expected lines that ruling party politicians make favorable statements for the policies, while the opposition is less positive about them. These statements by politicians on either side are highly polar because the politicians tend to either engage in blame-games with one another regarding the policy issues, or use rhetorical arguments about the promise of change through technology. Interestingly, we also find that prominent business-persons belonging to large business conglomerates tend to have favorable sentiments about the policies. Both politicians and business-persons alike generally talk about the benefits of these technological policies in solving issues of the poor. This indicates an alignment between business-persons and politicians, of a belief in high-modernism and technology determinism. On the contrary, we observe insignificant coverage of the

failures and loopholes of these technologies and ICTD policies, from the mass media, policymakers, and business-persons.

This high-modernistic discourse remains prevalent despite the ever increasing gap between its supply-side acceptance and demand-side acceptance [152], and failure of technological interventions like e-governance due to a supply-driven focus [98]. On similar lines, Solomon et al. [30] find that the digitization of land records under the ambitious *Bhoomi* scheme of e-governance led to increased corruption, and significantly increased time taken for land transactions. Pal [151] emphasizes that irrespective of the push by the government towards adopting newer technologies, technological innovations are a very small part of the challenge of social inclusion, and that the quest for solutions that yield more inclusive societies does not lie in the technology, but on the evolution of public sphere and its will to be more inclusive.

It is interesting to identify the reasons behind the aforementioned high-modernistic approach of the state, and its alignment with the viewpoints of the business-persons. Corporate-government interlocks can be one of the important reasons behind the alignment of high-modernistic views of politicians and business-persons. Such interlocks may cause a bidirectional flow of favors from the state to the corporations and vice versa, leading to their alignment of views. However, another important reason might be the need to survive – from Scott's and Harvey's arguments, we can see that both politicians and business-persons need to project technology as the ultimate solution to all problems primarily for their own respective survivals. Politicians constantly need to find new ways to improve lives of the citizens. The novelty of newer technologies is a low hanging fruit that could be made to pass as the 'final solution' to all problems, when detailed analysis of the problems, nuanced planning, and long term solutions are difficult and time consuming to achieve. Businesses, on the other hand, also need to sell newer technologies, replace the older ones, and replace workers continually, to make capital returns and to stay in the competitive market. Thus, although there might not always exist tangible interlocks between politicians and business-persons, there are broader reasons that may lead to alignment of their viewpoints regarding technical advancements.

Our analysis also helps us comment on potential biases of the mass media itself. We find that there generally is little coverage of negative aspects about the policies, and views of experts and civil society representatives, in the mass media, while politicians and business-persons get most of the coverage in policy discussions. This raises concerns on whether the Indian mass media is making people aware about the technicalities of different policies and the problems associated with them, or simply serving as a theater for politicians and political debates, or covering business-persons' views that are mostly tech-deterministic.

Siebert defines the four theories of press [189] namely, the *Authoritarian*, the *Soviet Communist*, the *Libertarian*, the *Social Responsibility* theories. According to the Authoritarian theory, the press and all of the information contained in it is controlled by the state or the government. The Soviet Communist theory provides higher control of the state on the press or media. According to this theory, not only does the government control the media or the information present in the media, but it also runs the media as a tool for its own propaganda. Libertarian theory stands opposite to the Authoritarian theory, and keeps the press or media out of state control. According to this theory, the primary duty of the press is to serve the interest of citizens by presenting the truth. Since a capitalistic society allows for free enterprises and corporate control of press or media, Social Responsibility theory states that the press should be made strong enough to function outside of any influence, be it corporate influence or state influence. We find from the last two theories that the role of media is to analyze and critique the state's policies, acting as a watch dog. Moreover, apart from informing, entertaining, and acting as a watch dog, media also has the responsibility of raising conflicts to the plane of discussion. Our findings point towards the contrary. In the next chapter, we study if there exist bias of any form in the Indian mass media, with respect to the content it presents on the policy issues.

Chapter 8

Towards a Fairness and Diversity Guaranteeing News Aggregator

Biases of different forms exist in the production, presentation, and distribution of news in online mass media outlets and in web data [22].These biases can occur in terms of coverage, selection, and sentiment slant [175]. Earlier works have shown that such biases exist in the news content produced by various Indian news-sources [182], among which one of the biases is the bias in terms of the coverage given to various topics (aspects) corresponding to a policy event. Although news recommendation algorithms can counter this bias in content by providing equitable attention to all of the topics relevant to a policy, several researches have shown that these algorithms themselves suffer from different kinds of biases in content recommendation [21, 51]. In this chapter, we attempt to address the research question: *Can we produce a news-feed that is unbiased and fair, in terms of its representation of news?*. The aforementioned papers have argued for the need to consider a temporal dimension, although for different reasons like countering presentation bias. We argue on similar lines, by proposing a recommendation algorithm that ensures fairness and diversity in coverage of aspects of an event over multiple lists or news-feeds generated over time, while considering the temporal change in the production of aspects.

News recommendation algorithms are managed by platforms, and the platform managers may want to define a clear editorial policy that they follow for selection of the news feed. We call a news recommendation algorithm *fair* if it provides equitable coverage to all aspects in a policy event over a period of time (i.e., across multiple news-feeds). While fairness ensures an equitable aspect coverage over a longer period of time across all news-feeds generated during that period, it is also important to ensure that over a shorter period, there is variety in the recently displayed aspects in the news-feeds generated. This phenomenon is called *diversity*. Long term fairness coupled with short term diversity ensures that every aspect obtains a fair chance to receive its share of exposure in the news-feeds generated by the algorithm, thereby countering the biases in recommendation of aspects, which often exist in existing news aggregators [42]. To ensure fairness (or diversity), a recommendation algorithm needs to follow a set of rules that assign weights to the news items or aspects, based on which they are displayed in its feeds. This set of rules is called a *fairness (or diversity) policy*, which platform managers can specify, and thereby convey a clear and transparent editorial policy to the platform users.

Our recommendation framework ensures long term fairness and short term diversity in the representation of news aspects, which we evaluate over four policy events, namely Demonetization [223], Aadhaar [226], GST [227], and Farmers' Protests [29] as reported in seven popular online Indian news-sources. As discussed in chapter 2, we have collected articles for these events, and formed a corpus also called the *media corpus*. We identify aspects from this media corpus considering a temporally evolving news-feed. Finally, to evaluate the performance of our algorithm, we compare it with the recommendation of Google Alerts (GA), along with two other baselines. Email based Google Alerts send news-emails on an event periodically based on certain keywords related to the event. We have collected alerts related to the four policies since February 2019 till August 2019 based on certain keywords. We study if there exists coverage bias in terms of the coverage given to the aspects in GA and the other baselines, and if so, if the recommendations provided by our framework is more equitable compared to them. We simulate live news recommendation by curating articles from the media corpus on a daily basis from a certain time period post the start date of an event, and evaluate our fairness and diversity

policies against our baselines. We find that our recommendation algorithm significantly outperforms the baselines for the four events in terms of fairness and diversity when using an optimal set of parameters, and performs fairly well with respect to average age of the news articles shown in its news-feeds (also called recency). The proposed framework using this algorithm is generalizable enough to be used on any other dataset, even outside the policy domain.

8.1 Related Work

While biases exist in the content produced by news outlets and social media, web based search engines also suffer from algorithmic biases that amplify the content based biases that already exist [21]. There are several studies that discuss ways to counter algorithmic biases of recommendation algorithms. One of the seminal studies by Celis et al.[50] is closest to our work. In this paper, the authors study a variant of the traditional ranking problem in the presence of fairness or diversity constraints. They consider the value of placing an item in a particular ranking position, the collection of important attributes (such as gender, race, and political opinion) of the items, and a collection of fairness constraints to output a ranking, which maximizes the value while respecting the constraints. On similar lines, Zehlike et al. [233] define and solve the *fair top-k ranking problem*. In this paper, the authors develop a ranking algorithm to maximize utility, while maintaining group fairness constraints – the requirement that for any position in the ranked list, a minimum proportion must be maintained for the group that is underrepresented in the population. Kearns et al. [117] consider the problem of selecting a set of individuals with the maximum utility from a population with incomparable traits (for example, soccer players, teachers, etc.). They solve this problem under the constraint that the true within-group CDF values are not known for an individual, and can only be estimated from a finite pool of candidates.

While our work is motivated from these aforementioned studies, our data differs from that

used in these works. First, in our approach, the optimization in terms of fairness and diversity happens over multiple instances of lists (news-feeds) being generated over time, whereas the other approaches optimize over single instances of lists. Additionally, in our case, we define the utility based on exposure of *aspects* corresponding to a policy event, alongside an item level utility based on the age (or recency) of articles. An aspect is a collection of many articles, and these articles within an aspect are prioritized based on their recency in our work. The utility varies across aspects for a policy event, and using a well defined utility function, we ensure that the aspect representation in our news-feed achieves fairness over long term and diversity over short term, while also ensuring recency of articles displayed for an aspect.

There also have been studies on designing systems to counter the biases existing in online news data. Park et al. [156] develop a novel system named *NewsCube*, which automatically detects topics corresponding to events in popular news, and provides a balanced viewpoint to the user by ensuring plurality in the topics displayed. On top of balancing news presentation, it also provides the users with a system to automatically tune their news reading habits. On similar lines, Munson et al. [139] design and deploy a browser widget designed to nudge its users to read balanced political viewpoints in the US scenario. This widget records the aggregate political lean of users' weekly and all time reading behaviors, in order to encourage the users to balance their news reading habits. Park et al. [157] design a framework to perform aspect (viewpoint) based classification of news articles, i.e., classification of news articles belonging to an event based on the different viewpoints from which the articles are written. This framework not only allows the users to go through different articles for an event, but also allows them to form their own, independent viewpoint based on a deep analysis and fine-tuning of the various aspects with which the articles are already written.

While most of these studies are based in the US political scenario where the political leaning or bias can be bipolar, we identify biases in online news data based on multiple aspects, constituencies, and political parties in India, and ensure fairness and diversity of news presentation with respect to the aspects. The aspect analysis approach used by Park

et al. [157] is to an extent similar to our aspect extraction. However, the difference lies in the computational approach – while they use the structure of news articles and evolution of their frames of presentation to capture aspects, we use LDA to identify aspects.

8.2 Data

We perform our analysis on 22302 articles on Demonetization (Nov 2016 to Dec 2019), 13908 articles on Aadhaar (Jan 2011 to Aug 2019), 22179 articles on GST (Jan 2011 to Dec 2019), and 85486 articles on Farmers' Protest (Jan 2011 to Nov 2019). These articles are collected using a keywords based approach as discussed in chapter 2.

We also collect email based Google Alerts data for the four policies, using the same keywords described in table 2.1, from February 2019 till August 2019. After removing duplicates and weeding out irrelevant articles, we finally have a total of 8191 articles for the four events. The final corpus contains 1392 articles for Demonetization, 1068 articles for Aadhaar, 2566 articles for GST, and 3165 articles for Farmers' Protests. Google Alerts is used as a baseline in our work and we compare the performance of our algorithm in terms of fairness and diversity, with that of Google Alerts, with respect to the relative coverage (as defined in chapter 5) provided in the news-feeds to various aspects in a policy.

The purpose of our news recommendation framework is to produce a fair and diverse news-feed that aids in countering the biases in aspect coverage, thereby achieving equity in aspect representation. To understand how fair or diverse our framework is, we compare it with Google Alerts, which is also used as a news recommendation system. To see if there exists biases in Google Alerts in terms of the coverage provided to the aspects belonging to a policy, we measure the relative coverage of each aspect corresponding to the four economic policies for Google Alerts. Our method of measuring the relative aspect coverage in mass media and in GA is the same as described in the study by Sen et al. [182]. Figure 8.1 shows the aspect coverage for both Google Alerts and the mass media.

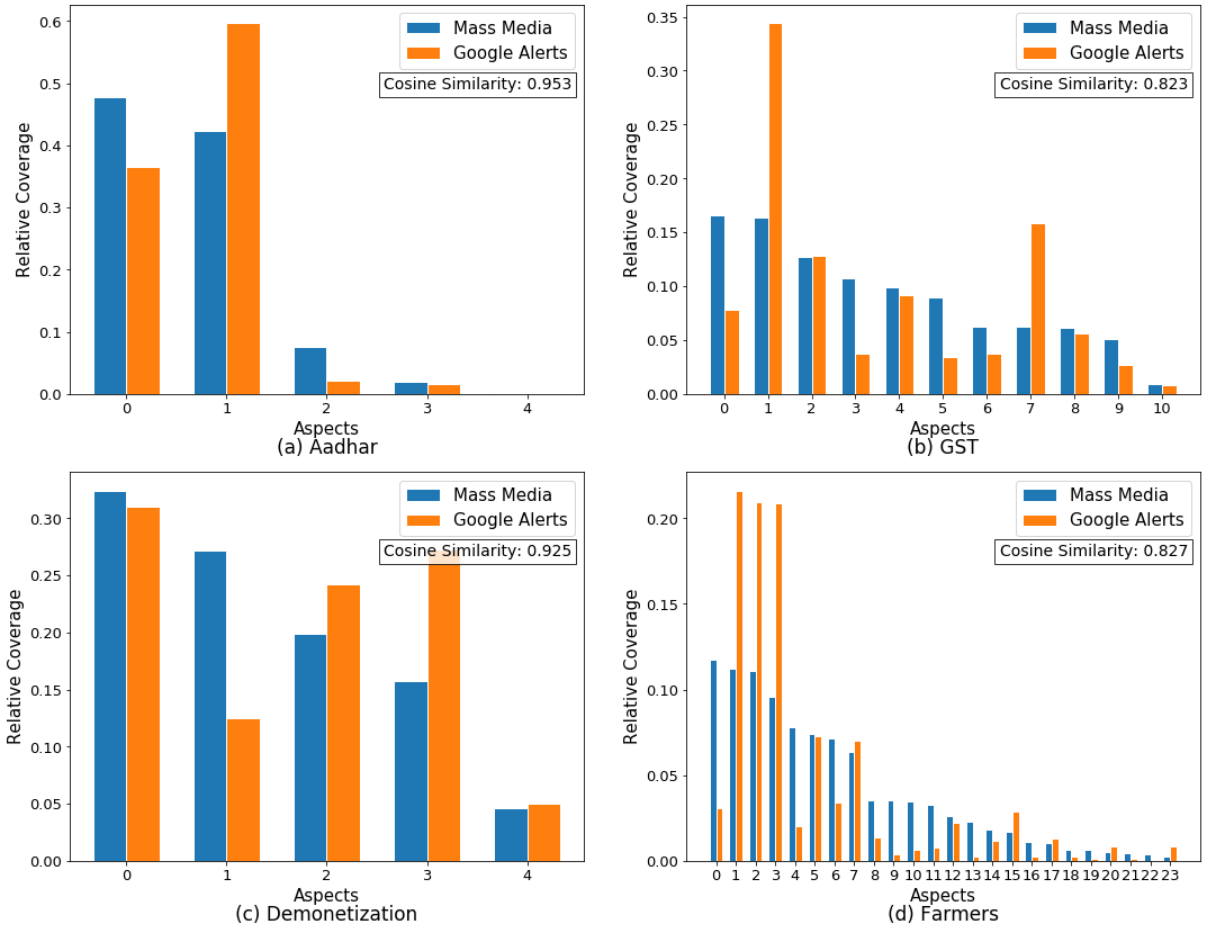


Figure 8.1: Comparison of the relative aspect coverage of Google Alerts and that of news-sources, showing a strong similarity in the aspect coverage trend followed by the two (cosine similarities indicated in boxes within the plots). The bias in aspect coverage is also evident in both of the sources. Further, Pearson Coefficient for the four cases are 0.92, 0.6, 0.6, 0.76, respectively.

From the plots, we find that the aspect coverage is significantly non-uniform for Google Alerts, which indicates the presence of bias in aspect coverage. This trend is similar to that observed for mass media as well. Additionally, we see that many of the top covered aspects in Google Alerts are also the top covered aspects in the Indian mass media, and the relative coverage trends of the two sources are significantly similar, as seen from the high values of cosine similarity between the aspect coverage. Thus, our plots indicate that GA does not sufficiently counter the biases that exist in aspect coverage, and that its bias is also significantly similar to the biases observed in mass media. This is evident from the high values of correlation between the aspect coverage of GA and mass media (0.92, 0.6, 0.6, 0.76 for Aadhaar, GST, Demonetization, and Farmers' Protests). Given these findings, we intend to develop a recommendation algorithm that counters such biases in aspect coverage, ensuring fairness and diversity in the recommendation of news aspects. In the following sections, we elaborate the architecture of our framework and the methodology followed behind the development of our recommendation algorithm.

8.3 System Architecture

The basic architecture of our framework is described in figure 8.2. Our framework consists of three broad stages of operation: **(A) Data Extraction**: The data extraction module crawls article data on a daily basis from the seven news-sources that we consider in this study using their RSS feeds. These articles are stored in an *article database* (also called the *media corpus*). We extract news articles related to a policy event from this article database next, using event specific keywords as discussed in chapter 2. The policy specific articles are used for further analysis in this work. **(B) Aspect Identification:** The aspect identification module identifies the aspects corresponding to the articles for an event, considering a temporally evolving news-feed. In a live set-up, for any policy event, new articles will be generated on a daily basis. Thus, we needed to develop an aspect identification module that dynamically analyses these articles periodically, and extracts the aspects relevant to them. **(C) News Recommendation:** This is the final

stage where our algorithm ensures fairness and diversity in news recommendation, and creates a daily news-feed. A list of these temporally evolving news-feeds are generated as output to the user. We do not consider the number of users who viewed an article in a feed, since we do not have that information. Our goal is to produce a news-feed everyday that ensures fairness and diversity in aspect representation. In the subsequent sections, we describe the aspect identification and news recommendation module.

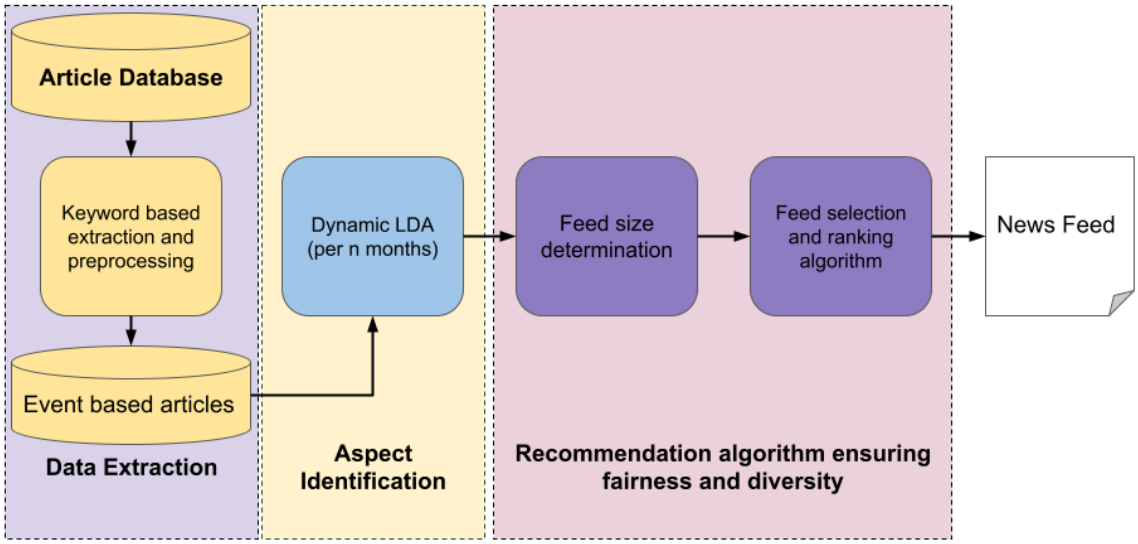


Figure 8.2: Architecture of our news recommendation framework

8.3.1 Aspect Identification for a Temporally Evolving Feed

We use Latent Dirichlet Allocation (LDA) to identify such aspects within each policy event as discussed earlier. Further details about aspect identification can be found in our paper [182]. In this aforementioned paper and as also described in the earlier chapters, we worked with a set-up where all the articles corresponding to a policy event were available, in which case it is easy to identify aspects accurately. However, for this analysis, we need to identify aspects considering a live environment. In other words, we need an

aspect identification approach that identifies aspects dynamically on new articles that are received over time. Considering this temporally evolving set-up, we first generate a topic model using LDA on an initial corpus of articles (articles of first six months) for a policy where this initial corpus size is decided experimentally. Next, for each chunk of new articles that arrive over a period of time, we have two options for aspect detection.

The first approach, also called *inferencing*, maps the new chunks of articles to the fixed set of aspects generated from the initial corpus. In this approach, the aspects are fixed and do not evolve over time, and out-of-vocabulary words are not considered during the mapping. We evaluate the performance of inferencing by comparing it with a benchmark, also called the *gold model*. The gold model is an LDA model that is trained on the entire article corpus for a policy, and can be assumed to be the best model, since it contains all of the articles. The performance comparison between inferencing and the gold model is done using the *positive percentage*, which evaluates the percentage of article pairs placed similarly in the two schemes. The details of calculating the positive percentage is provided in the Appendix.

The second approach, *retraining*, involves completely retraining the LDA model at fixed intervals, with the entire set of articles received till that time. This approach re-builds the entire model from scratch after fixed intervals, and hence, captures the evolution of aspects – newer aspects may develop over time, and existing aspects may merge or become irrelevant. Similar to the previous case, we compare the performance of retraining with that of the gold model as well. While inferencing consumes lesser time than retraining, the latter should perform better in terms of accuracy. To resolve this trade-off between cost and accuracy, we compare the two approaches and describe the results in the Appendix.

8.3.2 Recommendation Algorithm to Ensure Fairness and Diversity

The final module of our framework, the recommendation module, outputs one news-feed per day to the user, while ensuring fairness and diversity in aspect coverage. This module has two primary components as shown in figure 8.2: to determine the daily news-feed size, and to select and rank articles in the feed.

We first want to estimate the daily news-feed size that we should generate on a particular day. The feed size should dynamically adapt to the volume of articles produced recently for an event in the mass media. However, the number of articles produced for an event may drastically vary each day – while on some days we may get only a few articles for an event, on some other days there might be a sudden spike in the event popularity. In order to closely follow the daily production of articles for an event, while also smooth the excessive variation each day, we implement a protocol similar to the TCP re-transmission timeout calculation [1]. The following equations describe our protocol of estimating the daily feed size for an event:

$$C_{smooth} = (1 - \alpha) * C_{smooth} + \alpha * C_{today} \tag{8.1}$$

$$Dev_{smooth} = (1 - \beta) * Dev_{smooth} + \beta * (C_{today} - C_{smooth}) \tag{8.2}$$

$$feed - size = C_{smooth} + \gamma * Dev_{smooth} \tag{8.3}$$

C_{today} denotes the number of articles produced for an event on a particular day, and the first equation models the running average of the count of articles produced per day (stored in C_{smooth}). α is the smoothening parameter. Equation 8.2 similarly models the running average of the deviation (Dev_{smooth}) of the number of articles produced for a day from C_{smooth}. Finally, the feed size is determined by adding the average deviation (multiplied by a factor of γ) to the average count of articles. α varies between 0 and 1, and we keep the value of γ as 2 as is kept in the original protocol by typical implementations.

Once the news-feed size is decided for the day, our recommendation algorithm generates the news-feed while ensuring fairness and diversity in aspect exposure. We consider that an aspect is *exposed* if its articles have been displayed in one or more news-feeds. Note that unlike other studies, our definition of exposure of an aspect does not consider how many people viewed a particular article from that aspect, the number of clicks made on an article link, or the reading time corresponding to an article, since we do not have information on these parameters.

The broad approach of our recommendation algorithm is as follows: we compute a desired distribution of aspect exposure based on a set of rules (a policy to ensure fairness), and then we track the achieved distribution with the objective to bring the achieved and desired distribution to convergence. We now define some of these parameters that are used in our recommendation algorithm to ensure fairness and diversity, followed by our fairness and diversity policies. Finally, we explain our recommendation algorithm.

- **Desired distribution (D):** The desired distribution is defined as the distribution of relative exposure of aspects desired over a certain period by a set of rules (or a *fairness policy*). This period of time, over which the desired distribution is calculated, is termed the *fairness window*. In our case, the fairness window is defined as three months experimentally. We describe the rationale behind keeping the fairness window of three months in the Appendix.

  To ensure fairness, our fairness policy attempts to attain the desired distribution of aspect exposure in the news-feeds over the fairness window.

- **Achieved distribution (A):** The achieved distribution is defined as the distribution of aspect exposure achieved by a recommendation algorithm across news-feeds created in the fairness window, according to its recommendation policy. In our case, this recommendation policy ensures fairness, diversity, and recency in aspect exposure. There often exists a trade-off between the desired exposure and the achieved exposure, and our algorithm attempts to get the achieved distribution A as close as possible to the desired distribution D during the fairness window, and a penalty (or

loss) of $D_j - A_j$ is calculated for each *aspect-j* based on how much the recommendation policy causes the algorithm to deviate from the desired distribution. This penalty, also called the *utility*, helps the algorithm decide which aspect to choose for exposure in the upcoming news-feeds. In other words, if $D_j - A_j$ is high for *aspect-j*, the algorithm should have a higher propensity of sampling articles from *aspect-j* for exposure in the next feed, so that the loss is minimal. We exemplify the evolution of news-feeds over time using the infographic in figure 8.3. It shows the change in the loss $D_j - A_j$ as a new feed is produced each day. We consider three aspects corresponding to a policy event in this example. We see in the figure that in Day-1, the feed selection algorithm selects most articles from $aspect-2$ for exposure, since its loss $D_j - A_j$ is maximum. This leads to an increase its achieved exposure, resulting in a drop in the loss. Similarly, $aspect-1$ is given the most exposure on Day-2. Finally, on the fourth day, $aspect-2$ is not exposed anymore, since its loss or utility is already zero for Day-3.

- **Production distribution (P)**: The production distribution or the long term production function provides us the distribution of the relative coverage of aspects as produced by the news-sources over the fairness window.

- **Short term production distribution (SP)**: This distribution captures the production distribution of aspects over a shorter period of time. We consider a window of 15 days as this period based on earlier studies [148], and, also since we find this period to be suitable to capture the popularity of a news aspect in our dataset.

Ensuring Fairness and Diversity

In this section, we formally define the notions of *fairness* and *diversity*, and discuss our policy to ensure fairness and diversity in aspect exposure, alongside the baselines that we compare it with. Fairness, according to our definition, ensures that considering a long period of time (the fairness window), each aspect achieves a fair share of exposure as defined by the *fairness policy*. This in turn results in overall fairness over the entire

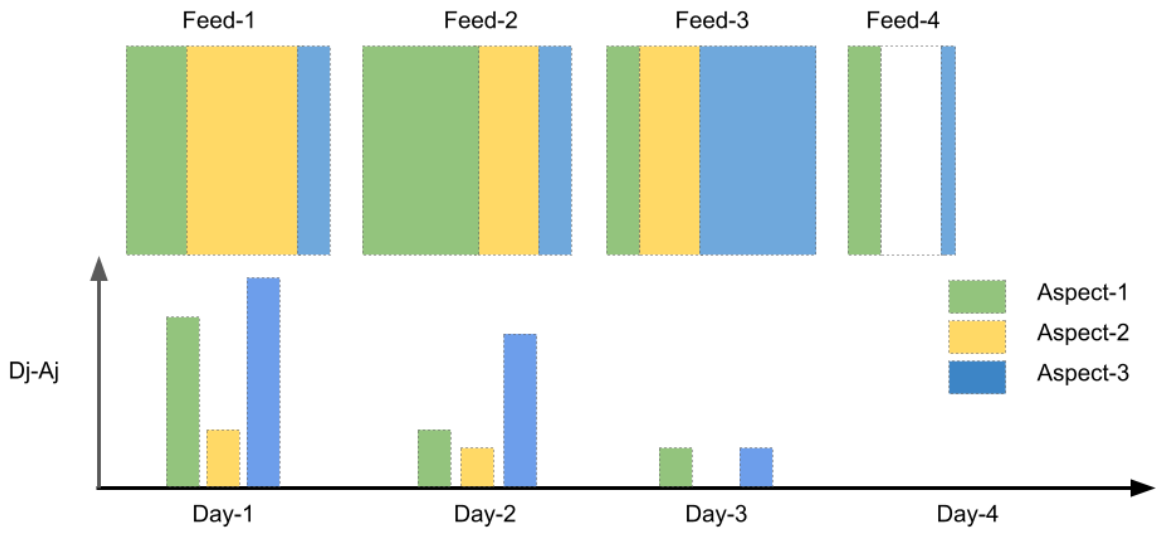


Figure 8.3: Evolution of news-feeds over time: we consider an event with three aspects using which daily feeds are produced by our algorithm. The aspects are selected for exposure in descending order of the corresponding loss as indicated by the width of the aspect in a feed (more the width, greater is the number of articles displayed from that aspect). Each time an aspect is exposed in a feed, its loss diminishes as A_j gets closer to D_j. The algorithm stops exposing an aspect when the loss is reduced to zero.

timeline of the event. Our algorithm ensures fairness on a rolling basis, i.e., the fairness window is calculated from the current date, and we consider all aspects (or articles) generated in the last three months from that date. There can be various ways or policies to define fairness. Each of these policies weighs the aspects corresponding to an event based on certain rules, and decides the exposure of each aspect in its news-feeds. We consider the following policies for study in this paper: **(a) Sampling according to production distribution:** In this policy, aspects are selected for exposure in a feed based on their production distribution during the fairness window. **(b) Latest news first:** This policy selects the k latest articles for exposure in the news-feed (of size k),

thus giving most importance to recency or age of articles exposed in a news-feed. (c) **Minimum threshold:** This policy ensures that over the fairness window, every aspect attains a certain minimum exposure, beyond which its exposure is decided based on the platform preference (production distribution P in our case). This policy ensures that no aspect has an exposure below the minimum threshold, while also providing importance to the platform manager's preference.

We use the last policy (minimum threshold) as the fairness policy that our algorithm implements. This is because it provides us a means to incorporate platform preference of aspect exposure, alongside ensuring a minimum threshold of aspect exposure. The other two policies are used as baselines for our work. We consider the production distribution P as the platform preference in our algorithm. The choice of P as the platform preference can be justified by the fact that for all of the considered events, P is significantly biased as can be assumed from figure 8.1. Thus, it can account for the biases which generally arise in platform preferences.

We assign a minimum threshold exposure of $f * 1/n$ to each aspect of an event according to our fairness policy where f is called the *fairness coefficient*. f can vary between 0 and 1. $f = 0$ indicates that the aspects will only be sampled based on their production distribution P, which is our first policy: *Sampling according to the production distribution.* On the other hand, $f = 1$ ensures equal coverage for all aspects. Any other value of f will translate to the *minimum threshold* policy where we give weightage to both the minimum threshold and the production distribution. Thus, we define our optimal minimum threshold fairness policy through equations 8.4 to 8.7:

$$D_j = f * \frac{1}{n} + (1 - f) * P_j \tag{8.4}$$

$$loss_j = D_j - A_j \tag{8.5}$$

$$score_j = \frac{loss_j - min(loss)}{max(loss) - min(loss)} \tag{8.6}$$

$$fscore_j = \frac{score_j}{\sum_j score_j} \tag{8.7}$$

In the above equations, D_j and A_j are as defined earlier. P_j is the relative production of *aspect-j* as defined by the production distribution in the fairness window. In equation 8.4, we linearly combine the two distributions as described above. We next calculate the loss or utility for *aspect-j* in equation 2, i.e., the difference between the desired and achieved exposure. $score_j$ captures the min-max normalized loss of *aspect-j*, and $fscore_j$ or the relative loss of *aspect-j* is the fairness score that is finally assigned to *aspect-j*. Our goal is to show that our recommendation approach with the minimum threshold fairness policy outperforms the baselines in terms of fairness, for an optimal value of $f = f*$. This optimal value should provide us the maximum fairness in aspect exposure. We discuss the calculation of $f*$ in section 8.4.

Diversity ensures that considering the aspects generated in a short period of time (in our case within the last 15 days from the current feed), there is sufficient variety in aspect representation in the current news-feed. We call this duration of 15 days the *diversity window*. The short term production distribution aids us in capturing the recently popular aspects while ensuring diversity. According to our diversity policy, we consider an equal relative coverage given to each aspect generated in the diversity window to ensure diversity. We define diversity with the following equation:

$$\begin{aligned} SD_j &= \frac{1}{n_d}, \text{ if } \textit{aspect-j} \text{ has at least one article generated in last 15 days} \\ &= 0, \text{ otherwise} \end{aligned} \tag{8.8}$$

where n_d is the number of aspects generated during the last 15 days, i,e., number of aspects which have at least one article generated during this window. SD_j defines the relative coverage of *aspect-j* under the current diversity policy.

Finally, we combine equations 8.7 and 8.8 linearly, to obtain the upper bound of the fraction of positions (U_j) in the news-feed that must be occupied by articles of *aspect-j*, for any position in the feed. In other words, while assigning an article from *aspect-j* at

any position k within the feed, *aspect-j* ideally should not occupy any more than $k * U_j$ positions in the top-k positions in the news-feed.

$$U_j = d * fscore_j + (1 - d) * SD_j \tag{8.9}$$

Here, d is termed as the *diversity coefficient*. Note that $d = 0$ ensures that the aspects are selected for the current news-feed only based on the diversity policy, while $d = 1$ ensures that the aspects are selected for the current news-feed only based on the fairness policy. Thus, d decides how much weightage we are assigning to fairness (and diversity). A high value of d will ensure more fairness in the exposure of aspects in the long term whereas a low value will focus on ensuring short term diversity. Thus, similar to the fairness coefficient, we also need to select an optimal value of $d = d*$, which ensures maximum diversity within feeds when compared to the baselines (while also retaining the fairness in aspect exposure). In other words, the pair $(f*, d*)$ should provide us the best performance in terms of both fairness and diversity. We discuss the calculation of both of these optimal parameters in section 8.4. Now that we have described the fairness and diversity policies, and the way our heuristic combines the two, we define our recommendation algorithm as described in Algorithm 1.

The algorithm does the following: we first identify the aspects (topics of discussion) for any event using the first six months of data available for the event to generate a news feed everyday during the timeline of the event (leaving aside the first six months). Next, in an online fashion, to produce a feed everyday, we first determine the $feed\_size$ (line 2), and consider articles published in mass media in the diversity window (last 15 days) for that event (line 3). This ensures recency of the articles selected for exposure. In lines 4 and 5, we calculate the upper bound U_j for each *aspect-j*. We start filling a feed by trying to pick articles for a feed-slot in the sorted order of when they were published, to ensure recency in our feeds to some extent. Thus, our algorithm selects an article from an aspect for exposure in a feed based on its age (utility at the item level). An article is filled in a slot if the upper bound constraint, $[exposure_j < k * U_j]$ is not violated. In case

Algorithm 1: Pseudo-code to ensure fairness, diversity, and recency across newsfeeds

Data: Aspects A identified using LDA within the first six months of news data available for an event

1 **foreach** $new\_day \in event\_timeline$ **do**
2 $feed\_size$ = Determine the size of feed from eq. 8.3;
3 $articles$ = Choose all articles belonging to this event published in the $diversity\_window$ (last 15 days), sorted in a reverse chronological order;
4 **foreach** $a_j \in A$ **do**
5 Calculate constraint U_j for aspect a_j using eq. 8.9;
6 Initialise $exposure_j = 0 \; \forall a_j \in A$;
7 **for** $k \leftarrow 1$ *to* $feed\_size$ **do**
8 $item$ = Pick the latest unpicked article from $articles$ if it belongs to an aspect a_j that does not violate the constraint U_j by checking $exposure_j$. In case no such article exists, pick the latest article that violates the constraint the least;
9 Place $item$ in the k^{th} position in the feed and output the feed;
10 $exposure_j = exposure_j + 1$;

no such article is found, the article whose aspect least violates the constraint is selected[1] from the last 15 days. After filling the articles, the algorithm updates the exposure of the aspects to which the articles belong (lines 7-10).

8.4 Results

In this section, we report the results with respect to the optimal values for the various parameters of the modules described in the previous section.

Aspect identification: We currently have two approaches that we can use on our policy data for aspect identification, namely inferencing and retraining. Generally, while

[1]This ensures that the feeds generated do not contain too many blank slots. We present a variation of this algorithm without constraint violation in the Appendix.

inferencing performs much faster compared to retraining as the corpus increases in size, retraining provides better accuracy in aspect identification. We experimentally decide using inferencing for our current study, and present these experiments in the Appendix.

While retraining the model might be the best option to ensure accuracy identification of aspects, in the interest of time, we stick to inferencing as the scheme of choice currently as it performs close to retraining for our dataset. In future, we intend to customize our algorithm to also take aspect evolution into consideration. We finally obtain 5 aspects for *Demonetization*, 24 aspects for *Farmers' Protest*, 11 aspects for *GST*, and 5 aspects for *Aadhar* using the first six months' data, and perform inferencing thereafter.

Feed size selection protocol: As discussed, we use a TCP re-transmission timeout based protocol to determine the size of the daily news-feed. The parameters α and β vary between 0 and 1. In order to see if the TCP based protocol provides us a daily feed size that approximately reflects the volume of articles produced each day, we plot the feed size and the number of articles per day for the four policies over time (we show these plots in the Appendix). We choose the parameters α and β to be 0.125 and 0.25, respectively, after experimentation. This helps us to achieve the right amount of smoothing of the feed size, while also gauging the interest for a particular event and reflecting that in the feed.

Choosing the optimal f and d values: To decide optimal values for the parameters f and d, we evaluate the performance of our recommendation algorithm on three metrics: (a) fairness, (b) diversity, and (c) recency, and compare it with the two baseline policies (the *Latest selection* and the *Sampling according to the production distribution* policies) and GA. We expect our algorithm (with its optimal parameters ($f = f*$) and $d = d*$ to perform better in terms of these metrics when compared to the baselines.

To evaluate fairness and diversity, we consider the relative distribution of aspects exposed in the news-feeds generated during the fairness and diversity windows respectively. For fairness, we calculate the GINI coefficient of this distribution within the fairness window, while to evaluate diversity, we calculate the Herfindahl Index (HHI) within the diversity window. These windows are rolled over the entire timeline in which we produce feeds,

and we obtain a GINI/HHI value for each day in this timeline. To measure recency, we calculate the age of a news-feed as the average age of the articles displayed in that feed. The age of an article is defined as the difference between the time of its exposure in the feed and the publication time of the article. Thus, if an article is displayed in a feed two days post its publication, its age is considered to be two days for that feed.

The advantage of using GINI coefficient for calculating fairness is that it reacts to zero values as well, unlike HHI. Thus, over the fairness window, if there are several aspects with zero exposure and some with high achieved exposures, it will indicate greater inequality and hence, the GINI would be high (indicating greater unfairness). On the other hand, diversity should ensure that within the diversity window, there is sufficient diversity of exposure among the available aspects. Hence, we use HHI for this purpose. As discussed, while f determines the weightage $(1 - f)$ given to the production distribution P, d determines the extent of overall fairness and diversity. To find out the optimal combination $P* = (f*, d*)$ that gives the best performance with respect to fairness, diversity, and recency, we define an optimization function F, which combines the three metrics. We perform a search on the (f, d) space to choose a region that minimizes F for all events. For a single f and d value, we obtain a series of daily GINI, HHI, and feed age (recency) values calculated over the entire timeline for which the feeds are produced as discussed previously. Median values from these series are selected to represent the fairness, diversity and recency components respectively. Once computed for the entire (f, d) space (for all (f,d) pairs), these components are separately normalized using min-max normalization. F then calculates the norm of these three components.

$$F = (fairness^2 + diversity^2 + recency^2)^{1/2} \tag{8.10}$$

Our goal is to see if there exists a point $P*$ in the (f, d) plane that minimizes F (i.e., maximizes the three metrics). Figure 8.4 shows the heat maps corresponding to the four policies. In these maps, the x-axis represents the diversity coefficient (d) and the y-axis represents the fairness coefficient (f). Each cell in the map corresponds to the resultant F, corresponding to a certain (f, d) combination.

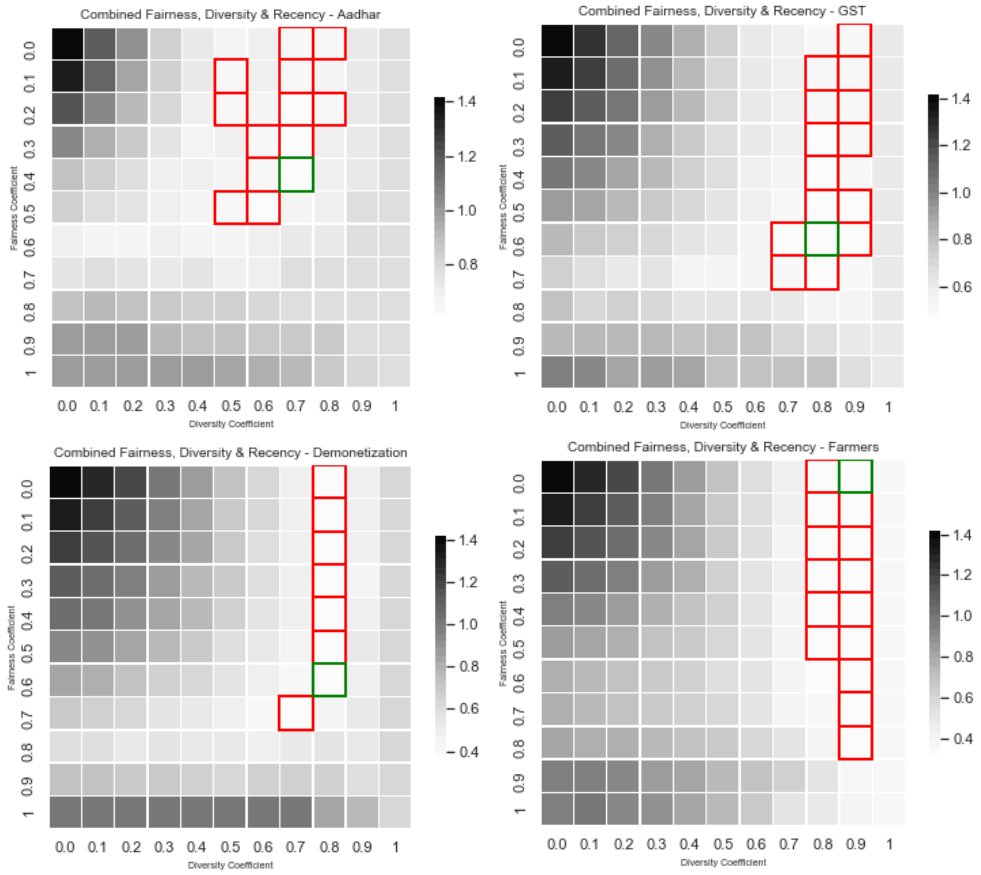


Figure 8.4: Heat map for combination of fairness, diversity, and recency corresponding to the four policies: the area with red borders indicate the zones where the algorithm performs decently in terms of fairness, diversity, and recency. The optimal values of fairness and diversity coefficients are chosen as (0.5, 0.8).

On studying these heatmaps, we find that there exists a region in the (f, d) space for which the algorithm performs well in terms of F, for all of the policy events (highlighted in red). There exists a trade-off between fairness/diversity and recency – while increasing f or d can lead to greater fairness and diversity, it leads to loss of recency. Hence, we choose an optimal value of $(f* = 0.5, d* = 0.8)$ in this region that performs well in terms

of all of the three metrics of fairness, diversity, and recency [2]. We choose this optimal $(f*, d*)$ pair for all further experiments in this chapter.

We now see how our algorithm performs with respect to the three baselines in terms of fairness, diversity, and recency. We show the GINI and HHI plots corresponding to Aadhaar and GST here. We also modify our approach to produce feeds containing aspects from multiple policy events, which we present in the Appendix. Our findings are consistent across all policies under consideration, and we present the other results in the Appendix. Figure 8.5 shows the GINI and HHI plots. The timeline considered for all of the plots presented hereafter is February 2019 to August 2019, since the GA data is available to us for this period. Our algorithm however outperforms the other two baselines for the entire event timeline as well. We find from these plots that our algorithm (with the optimal parameters of $f* = 0.5$ and $d* = 0.8$) performs better than all of the baselines and results in lower values of GINI and HHI coefficients for Aadhaar and GST. That is, in terms of both fairness and diversity, our algorithm outperforms the baselines. Additionally, we see that for the same set of optimal parameters, we could outperform the baselines in terms of fairness and diversity, for all of the policy events, and our recommendation framework could identify a common region (optimal regions highlighted in figure 8.4) in the (f,d) space for which it could obtain a fair and diverse feed for all of the events, which have significantly different production distributions (evident from figure 8.1). Therefore, our framework is able to ensure fairness and diversity across a diverse set of events using an optimal region in the (f,d) space, for a temporally changing production distribution. This is one of the contributions of our study.

We also analyzed if this difference in performance is significant by doing the Kolmogorov-Smirnov (KS) 2-sample test. The KS statistics for the four policies are: Aadhaar (GINI: 0.98, HHI: 0.88), Demonetization (GINI: 0.94, HHI: 0.87), GST (GINI: 0.94, HHI: 0.73), and Farmers' Protests (GINI: 1.0, HHI: 0.96). For all of the statistics reported, the p-value lies below 1e-7. This indicates that the difference in performance between our

[2]While the green grids in the maps show the most optimal (f,d) cells, to choose a common optimal value for all events, we allow a deviation of 10% from the best F value, and select $f* = 0.5$ and $d* = 0.8$.

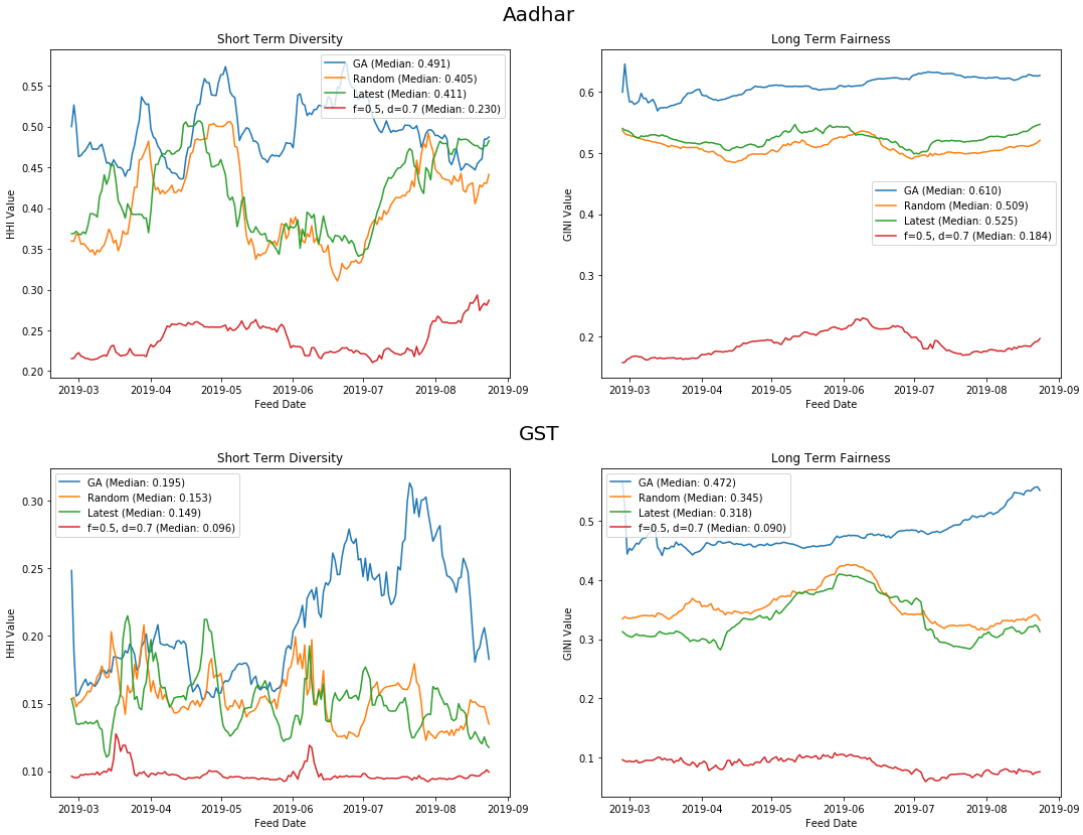


Figure 8.5: GINI and HHI Plots for Aadhaar and GST: our algorithm is seen to outperform all of the baselines for its optimal combination of parameters ($f* = 0.5, d* = 0.8$)

algorithm and the baselines is significant, in terms of fairness and diversity. We also find that Google Alerts performs the worst in terms of fairness and diversity in content presentation. These findings indicate that there is a greater need of self-regulation in terms of countering the biases existing in news aggregators similar to Google Alerts.

Measuring recency of news: While our algorithm with its optimal parameters aids in making the news feed fairer and more diverse when compared to the baselines, we also need to check if our generated news-feeds display the latest news to the users. This is important since exposing the latest news articles is a prime requirement for any news

aggregator. In our algorithm, we have tried to ensure recency of articles exposed by selecting the latest articles available for an aspect. To see how our algorithm with its optimal parameters fairs with respect to recency, we compare it with the three baselines as usual. We calculate the age of a news-feed as described earlier. We further average over a 15-day rolling window to remove noise, and plot the feed ages in figure 8.6.

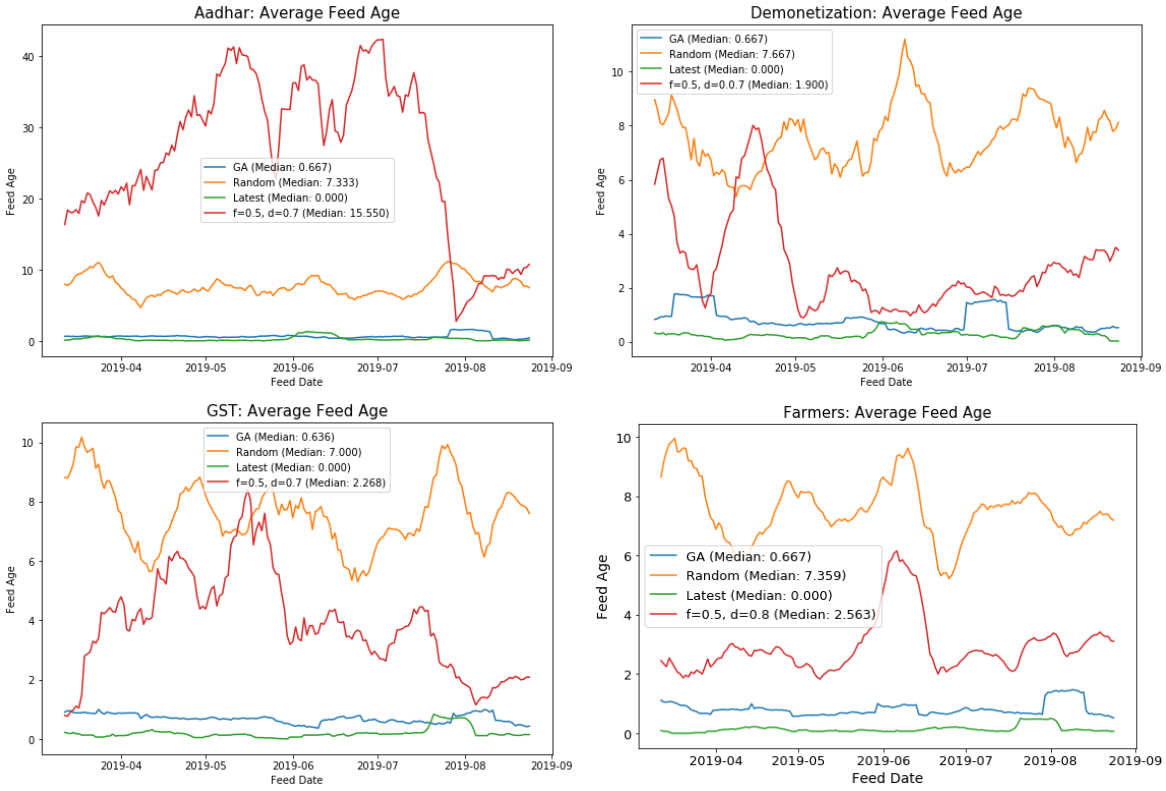


Figure 8.6: Weekly average news-feed age for our recommendation algorithm and the baselines

From these plots, we find that the *latest news first* policy outperforms all the other baselines, and our algorithm with its optimal parameters. This is because the *latest news first* policy works by selecting the latest news for exposure in the news-feed, and the only parameter that it ensures is the recency of articles. Next to it, Google Alerts performs

the best in terms of recency, since it sends updated news alerts on a daily basis over email. However, we see that our algorithm performs close to both of these baselines in most cases, and much better than the *Sampling according to production distribution* policy, which performs the worst. This loss in recency is due to the fact that the fairness and diversity constraints dominate our feed selection algorithm, and recency cannot be ensured in all cases considering our data. For example, if the desired exposure of an aspect is high, but no recent articles can be found for it, the algorithm ends up favoring old articles from the aspect. Such scenarios lead to loss of recency of our feed. As part of our future work, we will experiment further to ensure a better performance of our algorithm with respect to recency of the feeds. The mild loss in recency is, on the other hand, compensated by our algorithm's improved fairness and diversity where it performs much better than the baselines.

Measuring repetitiveness in feeds: It is at times difficult to find new articles for every feed produced by our algorithm, given its constraints. This situation arises when the algorithm has to expose an aspect (owing to the constraints) for which no new articles have been generated recently. This leads to repetition of already exposed articles. To measure repetitiveness, we calculate the proportion of articles that are repeated k times across all feeds produced by our algorithm, during the timeline of an event. We call this measure *repetition at k*. So, repetition at $k = 0$ would give us the proportion of articles that have never been shown for an event, $k = 1$ would mean the proportion of articles that have been shown only once across feeds, and so on. Figure 8.7 shows the comparison of our algorithm with the baselines in terms of its repetitiveness for Aadhaar and GST (rest of the event plots in the Appendix).

From the plots, it is clear there exists a trade-off between showing unique articles in different feeds, and repeating articles belonging to less represented aspects in the mass media to ensure fairness and diversity. The algorithm does not show quite a few articles in our feeds ($k = 0$). But we also see that the graph remains low for higher values of k. Therefore, the repetition allowed by our algorithm is moderate, and does not exceed a certain limit. We thus observe that our algorithm allows for moderate repetition of

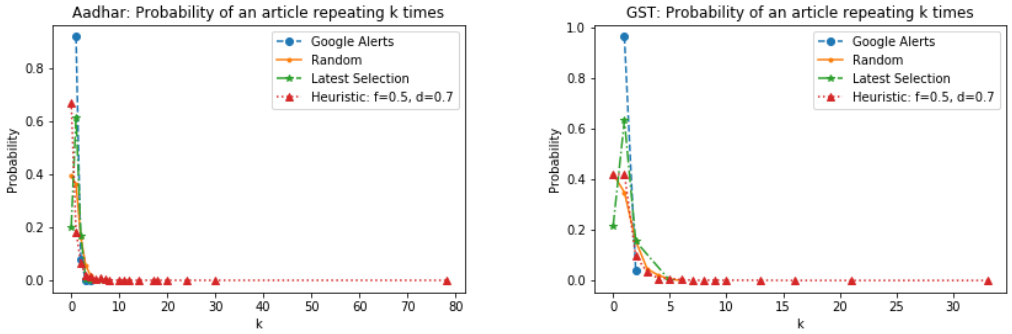


Figure 8.7: Repetition-at-k plots for the baselines and our algorithm, for all the four policies. *Note that for Google Alerts we do not plot for $k = 0$ as we do not have knowledge about the whole corpus of news articles from which it selects news. Thus, we do not know which articles it does not alert us about.*

articles in its feeds which could be an important design decision to be made by the platform managers.

8.5 Discussion and Conclusion

In this chapter, we attempted to answer the question: *Can we produce a news-feed that is unbiased and fair, in terms of its representation of news?*. We proposed a news-recommendation framework that attempts to ensure fairness, diversity, and recency in terms of aspect coverage, corresponding to four popular economic policy events, in this direction. Our findings prove that our recommendation framework successfully outperforms all of the baselines, including email based Google Alerts, in terms of fairness and diversity, thereby producing a series of fair and diverse news-feeds. Although there still lies some scope of improvement in terms of recency and repetition of articles presented in its feeds, our framework performs decently even on these metrics. The proposed framework with its recommendation algorithm is generalizable enough across multiple domains, and can be easily implemented on any dataset. Our analysis shows that considering the

current dataset of the four policy events studied in this chapter, our framework obtains an optimal value of the parameter pair (fairness and diversity coefficients) that ensures fairness and diversity across all of these events. These events have highly skewed aspect coverage as can be seen from figure 8.1, and hence, this guarantee of fairness and diversity using an optimal parameter combination is non-trivial. In the Appendix, we also show how our algorithm outperforms the baselines for multiple skewed production distributions that are synthetically generated on this dataset, thereby proving its robustness.

As discussed in the Related Work section, our work is similar to the work by Celis et al. [50]. In this work, the authors apply fairness and diversity constraints to the traditional ranking problem, and define the item-level utility based on attributes like gender, race, and political opinion relevant to the item. On the other hand, in our case, the recency of an article defines the item-level utility, i.e., our algorithm tries to rank articles based on their published dates, while maintaining the fairness and diversity constraints. An important difference between our work and those mentioned in the Related Work [50, 233] is that we consider the temporal evolution of aspects, while ensuring fairness and diversity on a list of news-feeds. The earlier studies mostly considered a single list of items while solving the ranking problem. The broad goal of our framework is to provide a balanced viewpoint about a policy event to the readers, which is similar to that of Park et al. [156]. However, we are yet to incorporate user level personalization in our framework.

There is a need of further research in the domain of fairness in recommendation algorithms as algorithmic biases amplify the already existing content biases in the data. Especially in the policy domain, these biases eventually result in a skewed perception of the public with respect to policy-making. In this direction, platform managers can use our framework to ensure greater fairness and diversity in content recommendation, and tune it according to their needs. Our work is a contribution towards self-regulation of recommendation platforms, which ensures equitability and diversification of content presentation.

Chapter 9

Conclusion and Discussion

In this thesis, we discussed our approach of analyzing the political economy around key policies in India. While political economy analysis encompasses a much broader area than what we include in this thesis, we focus on a few important facets that can be studied with the help of publicly available data. These facets are: (A) Analysis of corporate-government interlocks and their evolution over time, (B) Representation of some key policies in the three participants of democracy, namely the mass media, the social media, and the Parliament, and (C) Biases existing in these participants with respect to this representation. Finally, we discuss our initial work towards development of a fair news recommendation algorithm that attempts to counter this bias.

As discussed in the Related Work and in the previous chapters, there have been significant number of studies on political economy analysis. However, the contribution of this thesis lies in the fact that it proposes computer science based techniques for this analysis. We described the technological system that we have built using computer-based tools, which helps us in achieving this goal, using large-scale publicly available data. From our analysis, we found that the Indian mass media is biased, and that it shows preferential treatment towards certain constituencies and aspects, while nearly neglecting others. Specifically, immediate issues of the poor and technical nuances of problems in policy implementation

are often neglected. These biases are further echoed by the social media. Even in the Parliament, the policy discussion does not equitably cover issues of all sections of people, and the parliamentarians are seen to indulge in partisanship. Moreover, policy discourse in popular media highly covers the views of politicians and business-persons, and does not provide adequate attention to views of policy experts or academicians who can provide valuable insights on technical nuances for better policy-making.

As described in chapter 7, the media has a key responsibility of raising conflicts on policy events to the plane of discussion. With mass media and social media being one of the significant contributors of web data, the presence of biases suggests a stronger need for self-regulation of the Indian mass media, with respect to these pivotal roles. Further, the evident lack of coverage provided to the immediate issues of the poor enlarges the problem of digital divide. Our work serves as an initial step to address this issue by providing empirical justification of less representation to the issues of the poor, who might not have access to digital technology or social media. Our findings also suggest that although the policy discussions in the Parliament are nuanced, there is a need to focus more on the structural problems related to policies, and represent problems of all sections of people democratically. The Giant Economy Monitor website developed by us can help the policymakers and the target users in achieving this goal. Moreover, to ensure that the algorithms suggesting the news items to users do not reproduce these biases, we are currently building a recommendation system, which ensures long term fairness and short term diversity in representation of the various constituencies to which the news belongs. We believe that this platform and the techniques used can bring more visibility to the functioning of mass media, and push it closer to the goal of self-regulation through achieving diversity in content publication and educating the public of different viewpoints.

Our technological tool also helps us in finding evidence of increasing interlocks between corporate and government entities in India, from our analysis of publicly available web data. This increase indicates towards formation of a power structure, which can potentially influence policy-making, thus proving to be another important direction of political economy analysis. A majority of this increase can be attributed to the direct interlocks are

observed mostly between bureaucrats and firms, which indicate that bureaucrat-corporate interlocks might be one of the ways in which the corporate and government networks overlap. The other outcomes of corporate-government interlocks, apart from their influence on policy-making, have been studied in various works [72]. Our findings can further be used to red-flag such potential cases of exchange of favor or rent-seeking.

Overall, this thesis contributes towards the study of political economy through the use of computer science based tools and techniques, combined with qualitative content analysis. Broadly, we find that the three participants of democracy, namely the mass media, the social media, and the Parliament show indications of bias in policy representation, and are not representative enough to equitably cover the immediate concerns of all sections of people. In the next section, we discuss the challenges that we encountered while carrying out this research.

9.1 Primary Challenges

There were several challenges that we faced while conducting this research. The main technical challenges, which make the contribution in this thesis important, are listed as under:

Data collection: We showed that mass media and social media data, parliamentary questions asked on policy events, and corporate and political data are essential for the analysis of the political economy. It is important for such data should be in the public domain. Yet, we had to face several issues while collecting this data – we used standard crawlers and RSS feed parsers to collect article data, and several news-sources do not have archives of their old articles. Collecting historical data in such cases becomes difficult, since the structure of the websites of these sources keep changing frequently, requiring us to monitor the crawlers periodically, and change them as necessary. Similarly, several websites that we used to collect web based data changed over time (at times, even restricting data scraping), and we had to refer to multiple sources to collect the

same data. Additionally, collection of both media and web data at this scale was time consuming. Storage and backup of all this data was another challenge, since multiple copies of the same data needed to be retained to avoid data loss. We distributed our data across multiple servers to handle this issue. We therefore feel that publishing such data in a structured format needs an important consideration. Corporate information on Indian companies, for example, is present in a tabular format [144]. OpenCorporates [207] similarly hosts a knowledge base of corporations and their interconnections in a network format for different geographies. Similar structured data formats should be developed for disclosure of information present in the mass media and the Parliament.

Entity Resolution (ER): To have an in-depth understanding of the political economy, it is essential to have access to data on important people involved in the policy-making process. Thus, there exists a need of standardized disclosure of information about these important people. However, we found that the data collected from the wide variety of web based sources required entity resolution, since multiple entity names at times referred to the same person or object. Standard ER approaches did not suffice in this respect, since there does not exist any known approach that dynamically handles data of scale that compares to ours. Moreover, since we worked on Indian names, most of the existing resolution approaches on western names were found to be inapplicable to our domain. We found that resolving entities just based on their names and some commonly used information (like date of birth, designation, location, etc.) did not work well, in some cases due to the loss of accuracy, and in other cases simply because the information was not available in all of the sources considered. For instance, while we may have the date of birth of a politician in our knowledge base (KB), it might not be available in the mass media data where the same politician is being mentioned. We developed a context based ER approach, specifically customized to our data for this purpose. It uses context information like associated entities (in mass media data) and entity neighborhood (in corporate-government KB) for ER. Our infrastructure uses Mongodb, coupled with Elasticsearch based indexing to handle the problem of large-scale data analysis. There exist projects like GDELT [121] that curate news data from several sources, and identify influential people and organizations and present them in a standardized, network format.

Littlesis [105] shows the relationships between influential people and organizations in a network format similarly. The entities analyzed here include politicians, business-people, financiers, and their affiliated institutions. Standard knowledge bases like YAGO [203] contain knowledge of more than 10 million entities (like persons, organizations, and cities) curated from sources like Wikipedia and WordNet. We believe that our KB, alongside the analytical framework developed, can aid the users in gaining deeper insights on policy-making. Governments should disclose data on influential entities involved in policy-making in similar standardized formats.

Qualitative analysis: We use qualitative content analysis of data, along with computer science based techniques for political economy analysis. Qualitative analysis is bound to be subjective, and we had to keep checks at multiple levels to ensure minimum subjectivity. The first challenge was the preparation of the coding schema, which required us to manually peruse 100 articles from each policy event, and come up with relevant keywords, examples, and information that aided in accurately and unambiguously mapping each aspect to the five constituencies or frames of presentation. We went through multiple rounds of due-deliberation to finalize the schema. The second challenge was the aspect to constituency mapping exercise (using this schema). Despite making the schema as unambiguous as possible, coding errors and inconsistencies were obvious, since multiple annotators mapped the aspects of different policies to the constituencies. However, our schema ensured that the mapping exercise was fairly consistent as can be seen from the high values of inter-coder agreement.

Manual effort: Apart from the qualitative analysis, our work requires manually going through news articles at multiple phases of the study. For instance, we needed to go through articles of various aspects from a policy both for aspect naming and for measuring the accuracy of LDA in identifying the correct aspects. We had to perform manual analysis of articles also to measure the sentiment classification accuracy of the news articles, statements, and tweets on policies. These tasks were performed by multiple annotators, and inter-coder agreement was measured in most cases to ensure that there was not too much disagreement among the annotators. Including the qualitative analysis

part, we ensured that the data used for the manual tasks were of manageable size.

9.2 Limitations and Future Work

In this section, we discuss some of the limitations of this work, along with the future direction in which it can go in terms of its applicability. We discuss the limitations in terms of the three facets of political economy analysis that we study, and which we discussed in the beginning of this chapter. We also discuss some limitations of our current news recommendation approach to ensure fairness and diversity.

9.2.1 Analysis of Interlocks

One limitation of our analysis of corporate-government interlocks is the incompleteness and unavailability of data. Despite curating data from multiple sources, and trying a host of approaches for relationship extraction from media data, our KB is sparse in terms of its politician-company connections. This sparseness comes primarily from unavailability of data, since most politicians do not disclose their corporate connections publicly. We attempted to get these connections from alternative sources like the mass media articles. However, standard relationship extraction approaches like *Snowball* and *Association Rule Mining* did not fare well for such a large scale of data with complex sentence structures.

This also brings us to the various means of forming interlocks – currently, we only consider implicit (like location and family member based connections) and explicit connections (like board memberships) as interlocks, which are observable. However, several connections (especially between companies and politicians) are undeclared. In the Indian scenario, there exist several inactive *shell firms* using which financial transactions are maneuvered. Since there does not exist any known source of such data, we could not collect information on these interlocks. Also missing from our analysis is the consideration of other kinds of interlocks like *correlated movements* – for example, noticeable number of movements of

a pair of politician and bureaucrat across government departments, at multiple points in time. We intend to include these connections as well as part of our future work.

Another limitation of our approach would be the techniques that we use for our network computation. We use a PageRank based method to rank the entities based on various patterns of interlock. However, its comparison with other similar methods of ranking is currently missing. It must be stated though, that we have not come across many ranking approaches that apply on heterogeneous graphs (the corporate-government KB). The same applies to our method of calculating the indicator of corporate-government interlock.

Finally, while we compared our indicator of overlap with a random baseline, it would be important in future to validate this indicator based on feedback from experts and other sources of information (news articles, for instance) that provide us a nationwide picture of the state of corporate-government interlocks. We intend to do this validation in future.

Currently, our approach works on data existing in the public domain where evidence on interlocks is already available. As part of future work in this direction, we intend to extend our approach to also identify potential hot-spots or cases of interlocks. This will include identifying features corresponding to the entities, and building a model which will help us in red-flagging such potential cases. This thesis also does not touch upon the outcomes of the interlocks identified, i.e., the flow of favors that occur owing to these interlocks. It will be useful to study these outcomes by coupling the study with other information sources like news articles. Similar to the model intended to find potential cases of interlocks, one can also attempt to build a model to identify the use-cases where the already identified interlocks lead to illicit outcomes like rent-seeking.

9.2.2 Policy Representation and Bias Analysis

Currently our analysis of media bias towards or against prominent political parties is confined to the study of top 100 entities in these parties. We intend to improve this

in future by considering all entities relevant to a policy in a party, by consulting our corporate-government KB alongside mass media data. We are currently also improving our analysis of *by statements* (statements made by entities on policy events), to also include statements made by an entity regarding other entities and political parties. This can help in understanding pro/anti policy statements made in an indirect manner. For example, in an anti-policy article, an opposition member may make an adverse comment about the ruling party (or member of the ruling party), without mentioning the policy name at all in the statement. Currently, our approach does not capture such cases.

As discussed in the previous section, our work uses qualitative content analysis of data alongside computer science based techniques. Qualitative content analysis in our case involves building of a coding schema, and mapping of policy aspects to the five constituencies under consideration. Also integral to our work are multiple phases of manual analysis at different levels: aspect naming, accuracy checking of aspect classification, accuracy checking of sentiment analysis, and so on. These exercises are prone to subjectivity and manual errors. In future, crowd-sourcing and further fine-tuning of the qualitative and manual analysis approaches might help in addressing this issue.

An interesting direction for future work might also be the study of media effects. The different studies on media effects that happen in other geographies (for instance, the Pew Internet surveys [93] that happen in the US) do not happen often in India. However, such studies are much needed in the Indian context, given that we see evidences of media bias in India. As an example, if we consider the notion of relative coverage, it simply captures how much coverage the media house has provided to an aspect, relative to the other aspects for a policy event. While a skewed distribution of relative coverage indicates media bias, this information, when coupled with the information on the popularity of an outlet (both online and offline) can give us an idea of the impact that this bias can have – a biased newspaper with a large readership will have a much higher impact than one with a small reader-base. We have attempted to estimate the size of the online readership of Indian mass media outlets by collecting information on their follower communities on Twitter. However, this analysis can be broadened further by considering the offline

readership, and also by segregating the readership at a geographical level. The latter can help us analyze the impact of media bias on a local or regional level, and see if the location of readers is correlated with their perception of policy events in any way.

We also want to compare the coverage given by Indian mass media to various aspects, constituencies, and political parties with that given by international media. This will help us understand if the international media is also biased towards certain entities or issues, and whether these biases are similar to or different than the biases present in Indian mass media. Finally, another limitation of this work is that we study only English language news-sources in this thesis, since most natural language processing tools and techniques are applicable to English text. However, a small portion of the Indian population reads English newspapers. It is also important to capture the different aspects of political economy and media bias from vernacular news-sources, since many of these sources are widely read at a local level [104]. We intend to work in this direction in future.

9.2.3 News Recommendation

Among some of the limitations of our news recommendation framework, the consideration of user requirements is one. We have not deployed our framework yet, and hence do not possess data on user specifications of aspect coverage. A user study will enable us to understand if the proposed feeds by our recommendation algorithm is also considered fair by the users, and the extent to which it satisfies users. We intend to address this in future. As discussed earlier, there currently exists a trade-off between recency (and repetitiveness of articles) and fairness (plus diversity), according to our recommendation policy. This is another limitation of the current study, which we plan to mitigate by improving the algorithm with respect to the recency and repetitiveness of articles. Our framework should also consider aspect evolution, for which we need to periodically *retrain* our topic model to identify aspects, instead of the current *inferencing* process. This will also require us to modify our algorithm to dynamically map newer aspects to older aspects.

9.2.4 Reaching Out to the Users

The aforementioned sections discuss the challenges and limitations of our system, and the directions of future work that spawn from them. However, it is also indispensable to disseminate these information and findings of our framework to its intended users. This user base consists of journalists, policy researchers, social activists, economists, and the general public at large. We are currently in discussion with journalists from multiple media houses to obtain their valuable feedback on our system, and on the research questions asked by us regarding the political economy around policies. Their suggestions will be incorporated by the team that carries this work forward. As also discussed in chapter 1, we have built a website that contains the major findings from our research. This website is already live, and we will make it public soon. We believe that our system, when improvised with the suggestion of experts, can aid in further research on the political economy around key policies in any geography. It can also aid in making the people more aware of the political economy that affects their opinion, policies, democracy, and ultimately their lives and lives of others.

9.3 Recommendation Towards Accountability of Participants of Democracy

There are broadly two ways of ensuring accountability of the different participants of democracy, namely regulation and citizen led accountability. Our framework provides the users information relevant to the policy process on these participants. This is an important step towards regulating their functions. However, developing a sustainable process to ensure accountability of these participants needs existing institutions to be reformed or new institutions to be developed. We briefly discuss here some of our ideas in this regard.

9.3.1 Existing Bodies for Regulation

To ensure proper functioning of the various participants in a democracy, some form of regulation is required. Regulation can be ensured in two ways: (a) By monitoring effectiveness of the participant in terms of its roles and responsibilities, and (b) By holding participants accountable and taking appropriate punitive actions corresponding to violation of ethical codes and guidelines. In this section, we discuss the existing measures in both of these areas, and our suggestions in this regard.

Existing measures to ensure effectiveness: Mass media plays a significant role in influencing public opinion. Thus, the effectiveness of mass media in producing and disseminating information must be monitored. To ensure effectiveness of mass media, existence of a mass media watchdog or a media monitoring organization is important. There exists media watchdogs like The Hoot [5] in India, formed by practicing journalists to ensure media ethics, press freedom, and media's focus on development. Dissemination of fake news and misinformation also detrimentally affects mass media's credibility. In this direction, forums like Alt News [3] and Boom Live [4] debunk misinformation, and aid citizens in fact checking the news reaching them. There exist think tanks like the Pew Research Center [6] that conduct public surveys, demographic research, content analysis and other data-driven research to observe the trends followed in mainstream media, and the way people perceive them. We believe that such bodies that monitor the mass media on a regular basis can aid in moving the mass media closer to its responsibilities by better informing people, thereby making the media effective.

To ensure parliamentary effectiveness, it is essential to monitor the nature and type of discussions that occur in the Parliament, and provide appropriate feedback to the members when required. In this direction, organizations like PRS legislative research [8] attempt to make the Indian legislative process better informed, more transparent, and participatory. PRS interacts with members of the Parliament on a regular basis and provides them research inputs to support their work. It also engages with civil society organizations and the media to better inform them about policies and bills. The current

institutions to monitor the effectiveness of the Parliament and the mass media are civil society led, and are not institutionalized. While there can be different advantages and disadvantages of institutionalization, unbiased philanthropic funding will be important for the sustainability of these initiatives.

To monitor if the corporations are following best practices with respect to trade opportunities available to industries, statutory bodies like the Competition Commission of India (CCI) [7] can help. CCI aids in promoting competition among firms throughout India, and in preventing activities that have an adverse effect on competition like formation of cartels and rent seeking through political connections. Independent networks like International Consortium for Independent Journalists (ICIJ) [145] help in making public detailed information on powerful corporations, and the political and corporate elites connected to them. Similar data driven initiatives in India include How India Lives [2], which make corporate, government, and other information available to the people through detailed trend analysis and visualizations. Such information can help citizens keep an eye on the activities and correlated movements of corporations and their connections.

Most of the social media giants exercise self-regulation, and consist of internal regulatory boards that monitor and filter the content produced and spread through the media through moderation. While this can be effective in some cases, there are detailed discussions on their failures because of various reasons [154, 33]. Government intervention in monitoring and regulating the activities of social media might be helpful in this regard. For instance, post the data leakage fiasco involving Cambridge Analytica, the Indian government issued a notice to Facebook asking it for possibilities of data leakage of Indian citizens, and its influence on the Indian electoral process [198]. However, such steps might suffer from conflict of interest as has been recently reported [36, 159]. Civil society organizations can help in this respect, by forming bodies to monitor social media content through data scientific interventions, and making this information public to better inform people on misinformation or inappropriate content.

Existing measures to handle violation of standards: Some form of media regulation is present in almost every country, and Indian mass media is mostly self-regulated

[191]. The existing bodies for media regulation in India such as the *Press Council of India* and the *News Broadcasting Standards Authority* (NBSA) issue guidelines related to standard of media content and journalistic conduct [210]. However, these authorities cannot penalize media houses, their editors, and journalists for violation of these guidelines. Additionally, there are other limitations that restrict the operation of these regulators, e.g. the NBSA can only adjudicate complaints against 24 news broadcasters and their 64 channels, and not all of the 389 private news channels. Thus, more often than not, some media houses have been reported to flout the guidelines issued by these regulators [184]. The government of India has recently proposed the *Digital Media Regulation Bill* [109], which seeks to enforce registration of all digital media outlets with the Registrar of Newspapers for India. However, as argued by many experts [61], it fails to penalize media outlets in case of violation of the specified standards. Similar media regulation authorities exist in other parts of the world [229], e.g. the Press Recognition Panel in the UK, and the Federal Communications Commission in the US. These authorities have varying levels of control on media content.

Similar to media regulation, there also exists a need for regulation of the other participants of democracy. There are several reports on repeated disruptions of the Parliamentary sessions in India [11], which lead to loss of opportunity to hold the government accountable, and to deliberate on important legislative and policy issues. These disruptions also happen during the Question Hour, when the parliamentarians ask questions related to policy matters, which cover concerns of different sections of people. Another concern is the nature of discussions in the Parliament. Our work in this thesis showed how the Parliament at times fails to cover the structural issues related to the policies, and how some constituencies of people are covered by way of politicization. The latter is corroborated by studies like [114]. The *National Commission to Review the Working of the Constitution*, a regulatory body to recommend possible amendments to the Constitution, does provide pointers on the necessary attendance and number of sessions of the Parliament. However, this body does not possess any authority to enforce these recommendations directly or indirectly. There exists the parliamentary ethics committee to enforce and oversee the moral and ethical conduct of the members (MPs). While there exists a code of conduct

for MPs in the upper house of the Parliament, the committee has recently introduced the idea for a code of conduct for the lower house as well [127]. However, this committee cannot take note of the grievances of common people. Additionally, one of its major shortcomings is the failure to maintain the members' register of financial interest, and its lack of consideration for conflict of interests [165]. The speakers of the upper and lower houses of the Parliament also act as its regulators, but their power is limited to a few functions, like deciding the nature of the presented bills, and maintaining discipline and decorum of the house being some of them.

Finally, to handle violations of codes and ethics by any participant of democracy, reports on violations, or anomalies flagged by civil society should be made available transparently. Additionally, citizen awareness programmes should be organized on a regular basis to bring their attention to the various issues of democratic functioning.

9.3.2 Need for Citizen Led Accountability

There are several arguments in favor of self-regulation, like its efficiency, increased flexibility, increased incentives for compliance, and reduced cost. The self-regulating institution is supposed to have better knowledge of the sector to be regulated, and hence, is efficient in implementing the regulation process compared to a third party or the state. Self-regulation is said to be flexible, since the regulation process can be modified depending on the current circumstances of trade. Another advantage of self-regulation is that it encourages entities within the institution to respect and follow the regulatory norms, since they are formed by the institution itself. Finally, self-regulation is also less costlier, since the task of developing and overseeing the regulations is taken care of by the institution itself, instead of being routed through the government or another party.

There however exist several criticisms of self-regulation as well [48]. Some of them point out that self-regulation might encourage institutions to change the regulatory norms to unduly benefit themselves. Self-regulatory norms are generally subjective, which also

leads to their misuse. Additionally, self-regulation may also lead to rules, which ignore the views and concerns of the affected entities outside the institution. Self-regulation may also lead to anti-competitive conduct. Sometimes, institutions may flout their own regulatory rules, just to be at a competitive advantage. This also leads to the question of strength of enforcement of the regulations, and adequacy of the actions being taken against the dissenters.

Given these points, we believe that regulation can be made effective by striking the right balance between the autonomy given to an institution in terms of regulation, and the power vested in government bodies in enforcing them. On these lines, we believe that enforcement of the guidelines issued to mass media houses can be supervised by a government body. On similar lines, Douglas C. Michael [132] talks about "audited self-regulation", where a federal agency is given the responsibility of auditing the appropriate implementation and enforcement of regulatory norms. Additionally, civil societies and socially aware citizens can come together to form social media groups that inform the public about the conduct of these media houses with facts and figures. In this direction, fake news busting websites are also playing a significant role by countering misinformation spread by some media houses.

Since the Parliament is the highest decision making body, it also must be held accountable by the citizens whose well being depends on the policy discussions occurring in it. While the questions asked in the Indian parliament are regularly updated in the Parliament's website [9], they should be translated into vernacular languages, so that most people can understand the discourse. The ethics committee of the Parliament should also be allowed to consider a certain number of anonymous complaints and feedback from the citizens on parliamentary conduct and parliamentary discussions, on a regular basis. To avoid conflict of interest, the ethics committee should include eminent members from areas other than politics, like policy experts, academicians, and judiciary members. At the regional level, civil societies should participate to make the grievance reporting process against a member of the Parliament easier for citizens. They should also try to make sure that appropriate pressure is built on the authorities to work upon these grievances. One way

of doing this is by widely publicizing the issues in mass media and social media. Another way is to aid the citizens in gaining easy access to the opposition members, so that these grievances can also find their way to the Parliament. Finally, with the proliferation of web and IVR based communication systems, direct feedback of beneficiaries on the policy process should also be collected by the Parliament. This can be done through public polls, questionnaires, and surveys hosted on these platforms.

Bibliography

[1] Tcp timeout and retransmission, 2017. URL: http://isp.vsi.ru/library/Networking/TCPIPIllustrated/tcp_time.htm.

[2] How india lives, 2019. URL: https://howindialives.com/gram/.

[3] alt news, 2021. URL: https://www.altnews.in/.

[4] Boom live, 2021. URL: https://www.boomlive.in/.

[5] The hoot, 2021. URL: http://asu.thehoot.org/page/why-this-website-3.

[6] Pew research center, 2021. URL: https://www.pewresearch.org/.

[7] Competition commission of india, January 2021. URL: https://www.cci.gov.in/.

[8] Prs legislative research, January 2021. URL: https://www.prsindia.org/.

[9] Parliament of india lok sabha house of the people, Updated in 2020. URL: https://loksabha.nic.in/.

[10] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 u.s. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, pages 36–43, New York, NY, USA, 2005. ACM. URL: http://doi.acm.org/10.1145/1134271.1134277, doi:10.1145/1134271.1134277.

[11] admin_2. House of lost opportunities, 2013. URL: https://www.prsindia.org/hi/theprsblog/house-lost-opportunities.

[12] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM, 2000.

[13] Facundo Alvaredo, Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. World inequality report 2018. *The World Inequality Lab, http://wir2018. wid. world*, 2017.

[14] GINA ABENA Amedeka. *Newspaper Coverage of the 2010 District Assembly Election in Ghana: A Content Analysis of Daily Graphic and Daily Guide*. PhD thesis, University of Ghana, 2015.

[15] Jisun An, Meeyoung Cha, P Krishna Gummadi, and Jon Crowcroft. Media landscape in twitter: A world of new conventions and political diversity. In *ICWSM*, 2011.

[16] Jisun An, Daniele Quercia, and Jon Crowcroft. Partisan sharing: Facebook evidence and societal consequences. In *Proceedings of the second ACM conference on Online social networks*, pages 13–24. ACM, 2014.

[17] Adrian U Ang, Shlomi Dinar, and Russell E Lucas. Protests by the young and digitally restless: The means, motives, and opportunities of anti-government demonstrations. *Information, Communication & Society*, 17(10):1228–1249, 2014.

[18] Amelia Arsenault and Manuel Castells. Switching power: Rupert murdoch and the global business of media politics: A sociological analysis. *International Sociology*, 23(4):488–513, 2008.

[19] Amelia H Arsenault and Manuel Castells. The structure and dynamics of global multi-media business networks. *International Journal of Communication*, 2:43, 2008.

[20] Srikrishna Ayyangar and Suraj Jacob. Question hour activity and party behaviour in india. *The Journal of Legislative Studies*, 21(2):232–249, 2015.

[21] Ricardo Baeza-Yates. Data and algorithmic bias in the web. In *Proceedings of the 8th ACM Conference on Web Science*, pages 1–1. ACM, 2016.

[22] Ricardo Baeza-Yates. Bias on the web. *Communications of the ACM*, 61(6):54–61, June 2018.

[23] Sanjoy Bagchi. *The Changing Face of Bureaucracy: Fifty Years of the Indian Administrative Service*. Rupa, 2007.

[24] Ben H Bagdikian. Media monopoly. *The Blackwell Encyclopedia of Sociology*, 2007.

[25] Stefanie Bailer. People's voice or information pool? the role of, and reasons for, parliamentary questions in the swiss parliament. *The Journal of Legislative Studies*, 17(3):302–314, 2011.

[26] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[27] Abhijit Banerjee, Lakshmi Iyer, and Rohini Somanathan. History, social divisions, and public goods in rural india. *Journal of the European Economic Association*, 3(2-3):639–647, 2005.

[28] Abhijit Banerjee and Rohini Somanathan. The political economy of public goods: Some evidence from india. *Journal of development Economics*, 82(2):287–314, 2007.

[29] BBC. India farmers: Tens of thousands march against agrarian crisis, 30 November 2018. URL: https://www.bbc.com/news/world-asia-india-46396118.

[30] Solomon Benjamin, R Bhuvaneswari, and P Rajan. Manjunatha.(2007). bhoomi:'e-governance', or, an anti-politics machine necessary to globalize bangalore. *CASUM-m working paper*.

[31] W Lance Bennett and Shanto Iyengar. A new era of minimal effects? the changing foundations of political communication. *Journal of communication*, 58(4):707–731, 2008.

[32] Regina Berretta and Pablo Moscato. Cancer biomarker discovery: The entropic hallmark. *PloS one*, 5:e12262, 08 2010. doi:10.1371/journal.pone.0012262.

[33] Leonid Bershidsky. No, twitter, healthy conversation can't be engineered, March 2018. URL: https://www.bloomberg.com/opinion/articles/2018-03-02/twitter-s-attempt-at-self-regulation-won-t-work.

[34] Marianne Bertrand, Francis Kramarz, Antoinette Schoar, and David Thesmar. Politicians, firms and the political business cycle: evidence from france. *Unpublished working paper. University of Chicago*, 2006.

[35] Timothy Besley and Robin Burgess. The political economy of government responsiveness: Theory and evidence from india. *The quarterly journal of economics*, 117(4):1415–1451, 2002.

[36] Aurangzeb Naqshbandi Binayak Dasgupta and Sunetra Choudhury. Report on facebook's leniency to bjp members' communal posts causes row, August 2020. URL: https://www.hindustantimes.com/india-news/communal-postsreport-on-leniency-to-bjp-causes-row/story-43fWgMtLZQw1HUKlwnGPoM.html.

[37] Karen Bird. Gendering parliamentary questions. *The British Journal of Politics and International Relations*, 7(3):353–370, 2005.

[38] Kelly Blidook and Matthew Kerby. Constituency influence on 'constituency members': The adaptability of roles to electoral realities in the canadian case. *The Journal of Legislative Studies*, 17(3):327–339, 2011.

[39] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295, 2012.

[40] Thomas Bossuroy and Aline Coudouel. Recognizing and leveraging politics to expand and sustain social safety nets. 2018.

[41] Jules Boykoff. Framing dissent: Mass-media coverage of the global justice movement. *New Political Science*, 28(2):201–228, 2006.

[42] Engin Bozdag. Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15(3):209–227, 2013.

[43] Ceren Budak, Sharad Goel, and Justin M Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Quantifying Media Bias Through Crowdsourced Content Analysis (November 17, 2014)*, 2014.

[44] Ceren Budak, Sharad Goel, and Justin M Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271, 2016.

[45] ET Bureau. Ict spending in india will reach $144 billion in 2023: Globaldata, Jan 2020. URL: https://economictimes.indiatimes.com/tech/ites/ict-spending-in-india-will-reach-144-billion-in-2023-globaldata/articleshow/73138501.cms.

[46] Maintained by Ministry of Electronics and Information Technology. Cashless india, April 2020. URL: http://cashlessindia.gov.in/.

[47] Michael A Cacciatore, Ashley A Anderson, Doo-Hun Choi, Dominique Brossard, Dietram A Scheufele, Xuan Liang, Peter J Ladwig, Michael Xenos, and Anthony Dudo. Coverage of emerging technologies: A comparison between print and online media. *New media & society*, 14(6):1039–1059, 2012.

[48] Angela J Campbell. Self-regulation and the media. *Fed. Comm. LJ*, 51:711, 1998.

[49] Manuel Castells. *The network society A cross-cultural perspective*. Edward Elgar, 2004.

[50] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840*, 2017.

[51] Abhijnan Chakraborty, Saptarshi Ghosh, Niloy Ganguly, and Krishna P Gummadi. Can trending news stories create coverage bias? on the impact of high content churn in online news media. In *Computation and Journalism Symposium*, 2015.

[52] Carl R Chen, Yingqi Li, Danglun Luo, and Ting Zhang. Helping hands or grabbing hands? an analysis of political connections and firm value. *Journal of Banking & Finance*, 80:71–89, 2017.

[53] Chun-Fang Chiang and Brian Knight. Media bias and influence: Evidence from newspaper endorsements. *The Review of Economic Studies*, 78(3):795–820, 2011.

[54] Joshua Clinton, Simon Jackman, and Douglas Rivers. The statistical analysis of roll call data. *American Political Science Review*, 98(2):355–370, 2004.

[55] Joshua Cohen. Deliberation and democratic legitimacy. *1997*, pages 67–92, 1989.

[56] Maria M Correia. Political connections and sec enforcement. *Journal of Accounting and Economics*, 57(2-3):241–262, 2014.

[57] Marco Costantino, Richard G Morgan, Russell James Collingham, and R Carigliano. Natural language processing and information extraction: Qualitative analysis of financial news articles. In *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr)*, pages 116–122. IEEE, 1997.

[58] Pepper D Culpepper and Kathleen Thelen. Are we all amazon primed? consumers and the politics of platform power. *Comparative Political Studies*, page 0010414019852687, 2019.

[59] Ido Dagan, Lillian Lee, and Fernando CN Pereira. Similarity-based models of word cooccurrence probabilities. *Machine learning*, 34(1-3):43–69, 1999.

[60] Dave D'Alessio and Mike Allen. Media bias in presidential elections: A meta-analysis. *Journal of communication*, 50(4):133–156, 2000.

[61] Amrita Nayak Datta. Modi govt's new bill to regulate digital media too, has no punishment clause for paid news, November 2019. URL: https://theprint.in/india/modi-govt-new-bill-to-regulate-digital-media-has-no-punishment-clause-for-paid-news/326649/.

[62] Dries De Smet and Stijn Vanormelingen. The advertiser is mentioned twice. media bias in belgian newspapers. 2012.

[63] Bibek Debroy. *Indian Bureaucracy-Dismantling the Steel Frame*. Institute of South Asian Studies, National University of Singapore, 2008.

[64] Raj M Desai. The political economy of poverty reduction: Scaling up antipoverty programs in the developing world. *Wolfensohn Center for Development Working Paper*, (2), 2007.

[65] DHNS. Aadhaar may be used to create health record, 2013. URL: https://www.deccanherald.com/content/370507/aadhaar-may-used-create-health.html.

[66] DHNS. Rel retail mulls ecommerce push, 2015. URL: Readmoreat:https://www.deccanherald.com/content/483172/rel-retail-mulls-ecommerce-push.html.

[67] Nicholas Diakopoulos, Daniel Trielli, Jennifer Stark, and Sean Mussenden. I vote for—how search informs our choice of candidate. *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple, M. Moore and D. Tambini (Eds.)*, 22, 2018.

[68] Jean Drèze, Nazar Khalid, Reetika Khera, and Anmol Somanchi. Aadhaar and food security in jharkhand. *Economic & Political Weekly*, 52(50):51, 2017.

[69] David Edelmann. *Analysing and managing the political dynamics of sector reforms: a sourcebook on sector-level political economy approaches*. Citeseer, 2009.

[70] Peter Enderwick. What's bad about crony capitalism? *Asian Business & Management*, 4(2):117–132, 2005.

[71] Robert Epstein, Ronald E Robertson, David Lazer, and Christo Wilson. Suppressing the search engine manipulation effect (seme). *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22, 2017.

[72] Mara Faccio. Politically connected firms. *The American economic review*, 96(1):369–386, 2006.

[73] Jessica T Feezell. Agenda setting through social media: The importance of incidental news exposure and social filtering in the digital era. *Political Research Quarterly*, 71(2):482–494, 2018.

[74] Echo E Fields. Qualitative content analysis of television news: Systematic techniques. *Qualitative Sociology*, 11(3):183–193, 1988.

[75] Raymond Fisman. Estimating the value of political connections. *American Economic Review*, pages 1095–1102, 2001.

[76] Seth Flaxman, Sharad Goel, and Justin M Rao. Ideological and the effects of social media on news consumption. *Available at SSRN*, 2013.

[77] Florens Focke, Alexandra Niessen-Ruenzi, and Stefan Ruenzi. A friendly turn: Advertising bias in the news media. *Available at SSRN 2741613*, 2016.

[78] Department for International Development. Political economy analysis how to note, July 2009. URL: http://www.gsdrc.org/docs/open/po58.pdf.

[79] Susannah Fox. *The social life of health information 2011*. Pew Internet & American Life Project Washington, DC, 2011.

[80] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. The effect of collective attention on controversial debates on social media. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 43–52. ACM, 2017.

[81] Gina M Garramone and Charles K Atkin. Mass communication and political socialization: Specifying the effects. *Public Opinion Quarterly*, 50(1):76–86, 1986.

[82] Matthew Gentzkow and Jesse M Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

[83] Sean Gerrish and David M Blei. Predicting legislative roll calls from text. In *Proceedings of the 28th international conference on machine learning (icml-11)*, pages 489–496, 2011.

[84] CJ Hutto Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, 2014.

[85] Martin Gilens and Craig Hertzman. Corporate ownership and news bias: Newspaper coverage of the 1996 telecommunications act. *The Journal of Politics*, 62(2):369–386, 2000.

[86] Barney Glaser. *Experts versus Laymen: A Study of the Patsy and the Subcontractor*. Routledge, 2017.

[87] Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.

[88] Richard Gunther, Anthony Mughan, et al. *Democracy and the media: a comparative perspective*. Cambridge University Press, 2000.

[89] Amy Gutmann and Dennis Thompson. What deliberative democracy means. *Democracy: A Reader*, page 415, 2016.

[90] Jürgen Habermas. The structural transformation of the public sphere, trans. thomas burger. *Cambridge: MIT Press*, 85:85–92, 1989.

[91] Stephan Haggard, Robert R Kaufman, et al. *Development, democracy, and welfare states: Latin America, East Asia, and eastern Europe*. Princeton University Press, 2008.

[92] DP Hai. Process of public policy formulation in developing countries, 2016.

[93] Keith N Hampton, Harrison Rainie, Weixu Lu, Maria Dwyer, Inyoung Shin, and Kristen Purcell. *Social media and the'spiral of silence'*. PewResearchCenter, 2014.

[94] Barbara Harriss-White et al. *A political economy of agricultural markets in South India: masters of the countryside*. Sage Publications India Pvt Ltd, 1996.

[95] David Harvey. The fetish of technology: Causes and consequences. *Macalester International*, 13(1):7, 2003.

[96] David Harvey. *A brief history of neoliberalism*. Oxford University Press, USA, 2007.

[97] David Harvey. Neoliberalism is a political project. *Jacobin Magazine*, 2016.

[98] Richard Heeks. *Understanding e-governance for development*. Institute for Development Policy and Management Manchester, 2001.

[99] Edward S Herman. Manufacturing consent: The political economy of the mass media (2002, edward s. herman and noam chomsky; with a new introduction by the authors.; updated ed. of: Manufacturing consent. c1988.; includes bibliographical references and index. ed.), 1988.

[100] Alfred Hermida, Fred Fletcher, Darryl Korell, and Donna Logan. Share, like, recommend: Decoding the social media news consumer. *Journalism studies*, 13(5-6):815–824, 2012.

[101] The Hindu. Coal scam: Chronology of events, 2014 (updated June, 2016). URL: http://www.thehindu.com/news/national/coal-scam-chronology-of-events/article6350481.ece.

[102] Desheng Hu, Shan Jiang, Ronald E. Robertson, and Christo Wilson. Auditing the partisanship of google search snippets. In *The World Wide Web Conference*, pages 693–704, 2019.

[103] Xibing Huang, Dingtao Zhao, Colin G Brown, Yanrui Wu, and Scott A Waldron. Environmental issues and policy priorities in china: a content analysis of government documents. *China: An international journal*, 8(02):220–246, 2010.

[104] MRUC India. Indian readership survey 2019 q4, May 2020. URL: https://bestmediainfo.in/mailer/nl/nl/IRS-2019-Q4-Highlights.pdf.

[105] Public Accountability Initiative. Littlesis, Accessed on Jan 2018. URL: https://littlesis.org/.

[106] Lakshmi Iyer and Anandi Mani. Traveling agents: political change and bureaucratic turnover in india. *Review of Economics and Statistics*, 94(3):723–739, 2012.

[107] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideolog detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1113–1122, 2014.

[108] Craig Jeffrey and Jens Lerche. Stating the difference: State, discourse and class reproduction in uttar pradesh, india. *Development and Change*, 31(4):857–878, 2000.

[109] Lata Jha. Govt proposes bill to regulate digital media, November 2019. URL: https://www.livemint.com/politics/policy/govt-proposes-bill-to-regulate-digital-media-11574901878838.html.

[110] Simon Johnson and Todd Mitton. Cronyism and capital controls: evidence from malaysia. *Journal of financial economics*, 67(2):351–382, 2003.

[111] Ammu Joseph. Is the media watching poverty enough?, 2007. URL: http://www.indiatogether.org/medpov-op-ed.

[112] Florian Kaefer, Juliet Roper, and Paresha Sinha. A software-assisted qualitative content analysis of news articles: Examples and reflections. 2015.

[113] Bharathi Kamath. Intellectual capital disclosure in india: content analysis of "teck" firms. *Journal of Human Resource Costing & Accounting*, 2008.

[114] Devesh Kapur and Pratap Bhanu Mehta. *The Indian parliament as an institution of accountability*. United Nations Research Institute for Social Development Geneva, Switzerland, 2006.

[115] Devesh Kapur and Partha Mukhopadhyay. Sisyphean state: Why poverty programmes in india fail and yet persist. In *Annual Meeting of the American Political Science Association*, volume 30, 2007.

[116] Devesh Kapur and Milan Vaishnav. Quid pro quo: Builders, politicians, and election finance in india. *Center for Global Development Working Paper*, (276), 2011.

[117] Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. Meritocratic fairness for cross-population selection. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1828–1836. JMLR. org, 2017.

[118] Stuti Khemani. *Political cycles in a developing economy: effect of elections in Indian states*. The World Bank, 2000.

[119] Gary King, Benjamin Schneer, and Ariel White. How the news media activate public expression and influence national agendas. *Science*, 358(6364):776–780, 2017.

[120] Raksha Kumar. India's media can't speak truth to power, 2019. URL: https://foreignpolicy.com/2019/08/02/indias-media-cant-speak-truth-to-power-modi-bjp-journalism/.

[121] Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone. In *ISA Annual Convention*. Citeseer, 2013.

[122] Michael Levien. The land question: special economic zones and the political economy of dispossession in india. *The Journal of Peasant Studies*, 39(3-4):933–969, 2012.

[123] Alexandra Mallett, Maya Jegen, Xavier D Philion, Ryan Reiber, and Daniel Rosenbloom. Smart grid framing through coverage in the canadian media: Technologies coupled with experiences. *Renewable and Sustainable Energy Reviews*, 82:1952–1960, 2018.

[124] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[125] Gary Marcus and Ernest Davis. Eight (no, nine!) problems with big data. *The New York Times*, 6(04):2014, 2014.

[126] Karl Marx. *Das Kapital: Critique of Political "O Economy. First Volume*, volume 612. Felix Meiner Verlag, 2018.

[127] Liz Mathew. Ethics panel set to form code of conduct for lok sabha mps, December 2019. URL: https://indianexpress.com/article/india/ethics-panel-set-to-form-code-of-conduct-for-lok-sabha-mps-6147466/.

[128] Claire McLoughlin. Political economy analysis: Topic guide. *GSDRC, University of Birmingham, Birmingham*, 2014.

[129] A Mejía Acosta and J Pettit. Practice guide: A combined approach to political economy and power analysis. *Work in Progress Paper, SDC-DLGN, Brighton: IDS*, 2013.

[130] Sharon Meraz. Is there an elite hold? traditional media to social media agenda setting influence in blog networks. *Journal of computer-mediated communication*, 14(3):682–707, 2009.

[131] Atif R Mian and Asim Ijaz Khwaja. Do lenders favor politically connected firms? rent provision in an emerging financial market. *Rent Provision in an Emerging Financial Market (December 2004)*, 2004.

[132] Douglas C Michael. Federal agency use of audited self-regulation as a regulatory technique. *Admin. L. Rev.*, 47:171, 1995.

[133] Charles Mills. Wright: The power elite. *New York*, 1956.

[134] Atul Kumar Mishra. Newspapers in india and their political ideologies, 2015. URL: https://rightlog.in/2015/07/newspapers-in-india-and-their-political-ideologies/.

[135] Prasanna Mohanty. Labour reforms: No one knows the size of india's informal workforce, not even the govt, 2019. URL: http://delivery.acm.org/10.1145/3210000/3209581/p54-baeza-yates.pdf?ip=115.97.131.182&id=3209581&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&__acm__=1549915037_a695eb7be552f3c0798bafed27b82ca7.

[136] Joy Moncrieffe and Cecilia Luttrell. An analytical framework for understanding the political economy of sectors and policy arenas. *ODI, London*, 2005.

[137] Sendhil Mullainathan and Andrei Shleifer. Media bias. Technical report, National Bureau of Economic Research, 2002.

[138] Densua Mumford and Torsten J Selck. New labour's ethical dimension: statistical trends in tony blair's foreign policy speeches. *The British Journal of Politics and International Relations*, 12(2):295–312, 2010.

[139] Sean A Munson, Stephanie Y Lee, and Paul Resnick. Encouraging reading of diverse political viewpoints with a browser widget. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[140] Harish V. Nair. Aadhaar hearing: Right to life of poor more important than elite class' privacy concerns, says centre, 2015. URL: https://www.indiatoday.in/mail-today/story/aadhaar-hearing-privacy-supreme-court-1026499-2017-07-27.

[141] Viet-An Nguyen, Jordan L Ying, and Philip Resnik. Lexical and hierarchical topic regression. In *Advances in neural information processing systems*, pages 1106–1114, 2013.

[142] Matthew C. Nisbet and Bruce V. Lewenstein. Biotechnology and the american media: The policy process and the elite press, 1970 to 1999. *Science Communication*, 23(4):359–391, 2002. arXiv:https://doi.org/10.1177/107554700202300401, doi:10.1177/107554700202300401.

[143] Pippa Norris et al. *Digital divide: Civic engagement, information poverty, and the Internet worldwide*. Cambridge University Press, 2001.

[144] Ministry of Corporate Affairs. Mca sector mapping. http://www.mca.gov.in/MCA21/dca/efiling/NIC-2004_detail_19jan2009.pdf, 2009. Accessed: 10/12/2015.

[145] The International Consortium of Investigative Journalists. Icij offshore leaks database, 2021. URL: https://offshoreleaks.icij.org/pages/about.

[146] Alexandra Olteanu, Carlos Castillo, Nicholas Diakopoulos, and Karl Aberer. Comparing events coverage in online news and social media: The case of climate change. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, number EPFL-CONF-211214, 2015.

[147] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[148] Jake O'Neill. How long does a news story last?, Feb 2019. URL: https://www.vuelio.com/uk/blog/how-long-does-a-news-story-last/.

[149] Claudia Orellana-Rodriguez, Derek Greene, and Mark T Keane. Spreading the news: how can journalists gain more engagement for their tweets? In *Proceedings of the 8th ACM Conference on Web Science*, pages 107–116. ACM, 2016.

[150] Diana Owen. The new media's role in politics, 2018.

[151] Joyojeet Pal. Access technologies and accessibility for inclusion. *The International Encyclopedia of Digital Communication and Society*, pages 1–7, 2015.

[152] Joyojeet Pal. The technological self in india: From tech-savvy farmers to a selfie-tweeting prime minister. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*, pages 1–13, 2017.

[153] Joyojeet Pal, Priyank Chandra, Vaishnav Kameswaran, Aakanksha Parameshwar, Sneha Joshi, and Aditya Johri. Digital payment and its discontents: Street shops and the indian government's push for cashless transactions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[154] A.S. Panneerselvan. Can facebook's self-regulation ensure information hygiene?, March 2020. URL: https://www.thehindu.com/opinion/Readers-Editor/can-facebooks-self-regulation-ensure-information-hygiene/article31609588.ece.

[155] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group , The, 2011.

[156] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. Newscube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 443–452, 2009.

[157] Souneil Park, SangJeong Lee, and Junehwa Song. Aspect-level news browsing: understanding news events from multiple viewpoints. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 41–50, 2010.

[158] J Parthasarathy and D Agarwal. 'for make in india to succeed, harness technology, foster innovation', 2016. URL: https://indianexpress.com/article/india/india-news-india/for-make-in-india-to-succeed-harness-technology-foster-innovation/.

[159] Billy Perrigo. Facebook's ties to india's ruling party complicate its fight against hate speech, August 2020. URL: https://time.com/5883993/india-facebook-hate-speech-bjp/.

[160] Roger Pierce. Using content analysis. *Research Methods in Politics*, 2008.

[161] Thomas Piketty. Capital in the 21st century. 2014.

[162] Ravi Shankar Prasad. Digital india comes of age: Under the modi government it is giving rise to employment, entrepreneurship and empowerment, 2018. URL: https://blogs.timesofindia.indiatimes.com/toi-edit-page/digital-india-comes-of-age-under-the-modi-government-it-is-giving-rise-to-employment-entrepreneurship-and-empowerment/.

[163] PTI. Government should focus on structural issues of agriculture sector: Report, 2017. URL: https://economictimes.indiatimes.com/news/economy/agriculture/government-should-focus-on-structural-issues-of-agriculture-sector-report/articleshow/59738007.cms?from=mdr.

[164] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. Echo chambers on facebook. 2016.

[165] Bhanupriya Rao. Lok sabha has a new standing committee on ethics. but does it have enough teeth?, 2015. URL: https://factly.in/parliamentary-ethics-committee-india-lok-sabha-has-new-standing-committee-on-ethics-but-does-it-have-enough-teeth/.

[166] Stephen D Reese, Tim P Vos, and Pamela J Shoemaker. Journalists as gatekeepers. In *The handbook of journalism studies*, pages 93–107. Routledge, 2009.

[167] Thomson Reuters. Open calais. http://www.opencalais.com/, Accessed on Jan 2018.

[168] Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):23, Jul 2016. doi: 10.1140/epjds/s13688-016-0085-1.

[169] Filipe N Ribeiro, Lucas Henriqueo, Fabricio Benevenutoo, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna P Gummadi. Media bias monitor: Quantifying biases of social media news outlets at large-scale. 2018.

[170] David Ricardo. *On the principles of political economy*. J. Murray London, 1821.

[171] Neil M Richards and Jonathan H King. Three paradoxes of big data. *Stan. L. Rev. Online*, 66:41, 2013.

[172] Ronald E Robertson, David Lazer, and Christo Wilson. Auditing the personalization and composition of politically-related search engine results pages. In *Proceedings of the 2018 World Wide Web Conference*, pages 955–965, 2018.

[173] Thomas Saalfeld. Parliamentary questions as instruments of substantive representation: Visible minorities in the uk house of commons, 2005–10. *The Journal of Legislative Studies*, 17(3):271–289, 2011.

[174] Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. Social media news communities: gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1679–1684. ACM, 2013.

[175] Diego Sáez-Trumper, Carlos Castillo, and Mounia Lalmas. Social media news communities: gatekeeping, coverage, and statement bias. In *CIKM*, 2013.

[176] Dietram A Scheufele and David Tewksbury. Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of communication*, 57(1):9–20, 2006.

[177] James C Scott. *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press, 1998.

[178] Maynard S Seider. American big business ideology: A content analysis of executive speeches. *American Sociological Review*, pages 802–815, 1974.

[179] Holli A Semetko and Patti M Valkenburg. Framing european politics: A content analysis of press and television news. *Journal of communication*, 50(2):93–109, 2000.

[180] Anirban Sen, Priya Chhillar, Pooja Aggarwal, Sravan Verma, Debanjan Ghatak, Priya Kumari, Manpreet Singh Agandh, Aditya Guru, and Aaditeshwar Seth. An attempt at using mass media data to analyze the political economy around some key ictd policies in india. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*, page 21. ACM, 2019.

[181] Anirban Sen, Debanjan Ghatak, Kapil Kumar, Gurjeet Khanuja, Deepak Bansal, Mehak Gupta, Kumari Rekha, Saloni Bhogale, Priyamvada Trivedi, and Aaditeshwar Seth. Studying the discourse on economic policies in india using mass media, social media, and the parliamentary question hour data. In *Proceedings of the 2Nd ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '19, pages 234–247, New York, NY, USA, 2019. ACM. URL: http://doi.acm.org/10.1145/3314344.3332489, doi:10.1145/3314344.3332489.

[182] Anirban Sen, Debanjan Ghatak, Kapil Kumar, Gurjeet Khanuja, Deepak Bansal, Mehak Gupta, Kumari Rekha, Saloni Bhogale, Priyamvada Trivedi, and Aaditeshwar Seth. Studying the discourse on economic policies in india using mass media, social media, and the parliamentary question hour data. In *Proceedings of the 2nd*

ACM SIGCAS Conference on Computing and Sustainable Societies, pages 234–247, 2019.

[183] Frans Sengers, Rob PJM Raven, and AHTM Van Venrooij. From riches to rags: Biofuels, media discourses, and resistance to sustainable energy technologies. *Energy Policy*, 38(9):5013–5027, 2010.

[184] Geeta Seshu. Cocking a snook at the nbsa-self regulation is not working, 2018. URL: http://asu.thehoot.org/media-watch/law-and-policy/cocking-a-snook-at-the-nbsa-self-regulation-is-not-working-10559.

[185] Aaditeshwar Seth and Jie Zhang. A social network based approach to personalized recommendation of participatory media content. In *ICWSM*, 2008.

[186] Dafna Shahaf and Carlos Guestrin. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–632. ACM, 2010.

[187] Sanjeev Sharma. Gst to hit consumers, unorganized jobs most, 2016. URL: https://www.tribuneindia.com/news/nation/gst-to-hit-consumers-unorganised-jobs-most/275659.html.

[188] Shantanu Nandan Sharma. Ias officers enjoy the freedom in the private sector, 2008. URL: https://economictimes.indiatimes.com/special-report/ias-officers-enjoy-the-freedom-in-the-private-sector/articleshow/2788619.cms?from=mdr.

[189] Fred Siebert, Theodore Bernard Peterson, Theodore Peterson, and Wilbur Schramm. *Four theories of the press: The authoritarian, libertarian, social responsibility, and Soviet communist concepts of what the press should be and do.* University of Illinois press, 1956.

[190] Yanchuan Sim, Brice DL Acree, Justin H Gross, and Noah A Smith. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101, 2013.

[191] Simran. Regulation of media in india - a brief overview, 2011. URL: https://www.prsindia.org/hi/theprsblog/regulation-media-india-brief-overview.

[192] Adam Smith and Dugald Stewart. *An Inquiry into the Nature and Causes of the Wealth of Nations*, volume 1. Wiley Online Library, 1963.

[193] Jackie Smith, John D McCarthy, Clark McPhail, and Boguslaw Augustyn. From protest to agenda building: Description bias in media coverage of protest events in washington, dc. *Social Forces*, 79(4):1397–1423, 2001.

[194] Katherine Clegg Smith and Melanie Wakefield. Textual analysis of tobacco editorials: How are key media gatekeepers framing the issues? *American Journal of Health Promotion*, 19(5):361–368, 2005.

[195] Katherine Clegg Smith, Melanie Wakefield, Catherine Siebel, Glen Szczypka, Sandy Slater, Yvonne Terry-McElrath, MSA Sherry Emery, and Frank J Chaloupka. Coding the news: the development of a methodological framework for coding and analyzing newspaper coverage of tobacco issues. *Impact Teen*, 2002.

[196] The Telegraph Special Correspondent. Scrap aadhaar in pds, demand food activists, 2017. URL: https://www.telegraphindia.com/states/jharkhand/scrap-aadhaar-in-pds-demand-food-activists/cid/1339184.

[197] Janaki Srinivasan and Aditya Johri. Creating machine readable men: legitimizing the'aadhaar'mega e-infrastructure project in india. In *Proceedings of the Sixth International Conference on Information and Communication Technologies and Development: Full Papers-Volume 1*, pages 101–112, 2013.

[198] Tech2 News Staff. Indian government issues notice to facebook asking for information on the cambridge analytica data breach, March 2018. URL: https://www.firstpost.com/tech/news-analysis/indian-government-issues-notice-to-facebook-asking-for-information-on-the-cambridge-analytica-data-breach-4410297.html.

[199] John Stasko, Carsten Görg, and Zhicheng Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, 7(2):118–132, 2008.

[200] Stephanie. Inter-rater reliability, 2016. URL: https://www.statisticshowto.datasciencecentral.com/inter-rater-reliability/.

[201] James Steuart. *An inquiry into the principles of political economy*, volume 2. 1767.

[202] Joseph E Stiglitz. *The price of inequality: How today's divided society endangers our future*. WW Norton & Company, 2012.

[203] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706, 2007.

[204] Sandip Sukhtankar. Sweetening the deal? political connections and sugar mills in india. *American Economic Journal: Applied Economics*, 4(3):43–63, 2012. URL: http://www.jstor.org/stable/23269730.

[205] Surabhi. 'aadhaar has no role in plugging leakages', 2018. URL: https://www.thehindubusinessline.com/economy/budget/aadhaar-has-no-role-in-plugging-leakages/article10024685.ece.

[206] JD Sutter. Texts, maps battle haiti cholera outbreak. *Retrieved October*, 31:2010, 2010.

[207] Chris Taggart and Rob McKinnon. Opencorporates: The largest open database of companies in the world, Accessed on Jan 2018. URL: https://opencorporates.com/.

[208] Elasticsearch Developers' Team. Elasticsearch, Accessed on Jan 2018. URL: https://www.elastic.co/.

[209] Neo4j Developers' Team. Neo4j, Accessed on June 2020. URL: https://neo4j.com/.

[210] NL Team. Nbsa directs aaj tak, zee news, india tv, news24 to apologise for violating ethics in sushant singh rajput coverage, 2020. URL: https://www.newslaundry.com/2020/10/08/nbsa-directs-aaj-tak-zee-news-india-tv-news24-to-apologise-for-violating-ethics-in-sushant-singh-rajput-coverage.

[211] Philip E Tetlock. Personality and isolationism: Content analysis of senatorial speeches. *Journal of Personality and Social Psychology*, 41(4):737, 1981.

[212] Paranjoy Guha Thakurta. Media ownership trends in india. *The Hoot*, 3, 2012.

[213] Paranjoy Guha Thakurta and Abir Dasgupta. An unnamed ias officer levels serious allegations of corruption against big-four firm kpmg india, 2018. URL: https://caravanmagazine.in/vantage/kpmg-unnamed-ias-officer-corruption-letter-modi.

[214] Kathleen Thelen. Regulating uber: The politics of the platform economy in europe and the united states. *Perspectives on Politics*, 16(4):938–953, 2018.

[215] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, December 2010. URL: http://dx.doi.org/10.1002/asi.v61:12, doi:10.1002/asi.v61:12.

[216] Manu P Toms. Happy govt is backing aadhaar, says nilekani, 2015. URL: https://www.hindustantimes.com/india/happy-govt-is-backing-aadhaar-says-nilekani/story-MJh9BHYyC6mL54XeYpVEaL.html.

[217] The Telegraph TT Bureau. 'digital india' has hurt middlemen, who are now spreading lies, says pm modi, 2018. URL: https://www.telegraphindia.com/india/lsquo-digital-india-rsquo-has-hurt-middlemen-who-are-now-spreading-lies-says-pm-modi/cid/1348276.

[218] Milan Vaishnav and Saksham Khosla. The indian administrative service meets big data. *Washington DC*, 2016.

[219] Peter Van Aelst and Stefaan Walgrave. Minimal or massive? the political agenda-setting power of the mass media according to different methods. *The International Journal of Press/Politics*, 16(3):295–313, 2011.

[220] Frank HM Verbeeten, Ramin Gamerschlag, and Klaus Möller. Are csr disclosures relevant for investors? empirical evidence from germany. *Management Decision*, 2016.

[221] Jessica S. Wallack. India's parliament as a representative institution. *India Review 7.2 (2008)*, 2008.

[222] Wikipedia. Digital divide. URL: https://en.wikipedia.org/wiki/Digital_divide.

[223] Wikipedia. 2016 indian banknote demonetisation, 2016. URL: https://en.wikipedia.org/wiki/2016_Indian_banknote_demonetisation.

[224] Wikipedia. Digital india, April 2020. URL: https://en.wikipedia.org/wiki/Digital_India.

[225] Wikipedia. National e-governance plan, April 2020. URL: https://en.wikipedia.org/wiki/NeGP.

[226] Wikipedia. Aadhaar, Last Modified in May 2020. URL: https://en.wikipedia.org/wiki/Aadhaar.

[227] Wikipedia. Goods and services tax (india), Last Modified in May 2020. URL: https://en.wikipedia.org/wiki/Goods_and_Services_Tax_(India).

[228] Wikipedia. 2g spectrum scam, Updated Jan 2018. URL: https://en.wikipedia.org/wiki/2G_spectrum_scam.

[229] Wikipedia. Media regulation, Updated on 2020. URL: https://en.wikipedia.org/wiki/Media_regulation.

[230] Kaibin Xu. Framing occupy wall street: a content analysis of the new york times and usa today. *International Journal of Communication*, 7:21, 2013.

[231] Shyamlal Yadav. Bureaucrats: Life begins at 60, 2011 (updated Jan 2018). URL: http://indiatoday.intoday.in/story/life+begins+at+60/1/125878.html.

[232] Sevgi Yigit-Sert, Ismail Sengor Altingovde, and Özgür Ulusoy. Towards detecting media bias by utilizing user comments. In *Proceedings of the 8th ACM Conference on Web Science*, pages 374–375. ACM, 2016.

[233] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578, 2017.

[234] Chengqi Zhang and Shichao Zhang. *Association rule mining: models and algorithms*. Springer-Verlag, 2002.

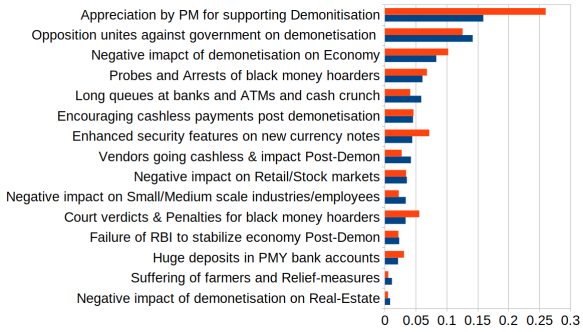
Appendices

Appendix A

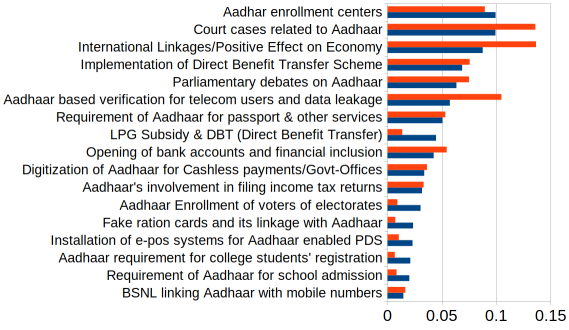
Analysis of Bias in Mass Media Content

A.1 Relative Coverage of Aspects

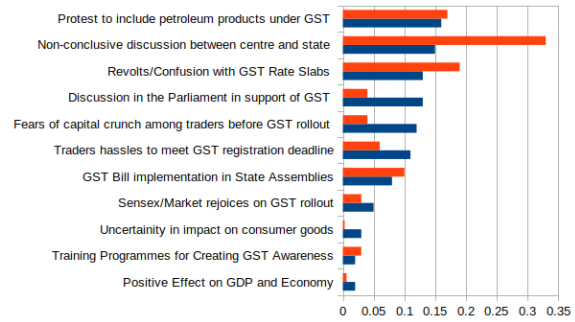
In this section, we present the relative aspect coverage provided by the mass media and social media to different policies, per aspect. We look at the aggregate coverage distributions here. Aggregate aspect coverage across all news-sources is calculated by summing up the number of words belonging to the aspect across these sources, and then dividing this number by the total number of words across all aspects in all news-sources for that policy. Figure A.1 shows the distributions. From the plot, it is evident that the social media follower community closely follows the coverage trend of aspects in the mass media. This is also established in our main paper. The correlation coefficients between the two vectors of aspect coverage are 0.92 for Demonetization, 0.90 for Aadhaar, 0.94 for GST, and 0.66 for Farmers'Protest. The low correlation for the last event arises from just two aspects: *Irrigation concerns and water pollution* and *Crimes and suicides in farmer community*. These two aspects are much highly posted about on Twitter (compared to the mass media), being sensitive issues related to the farmers. For the other three policies,



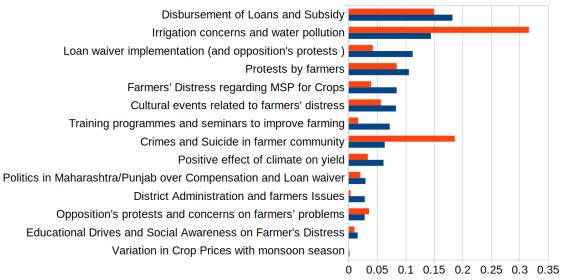
(a) Demonetization



(b) Aadhaar



(c) GST



(d) Farmers' Protest

Figure A.1: Aggregate relative coverage provided by the mass media and its social media follower community corresponding to each policy: the blue bars and red bars represent mass media and social media coverage, respectively.

the high values of correlation indicate towards a high alignment between the mass media and social media aspect coverage. We obtain similar trends even when we do this analysis for individual news-sources, which we do not report in this paper.

From the plots, it is also clear that both mass media and social media are biased in terms of aspect coverage, i.e., there exists a significant imbalance in the coverage of aspects. To state empirically, for Demonetization, Aadhaar, GST, and Farmers' Protests, the relative coverage in mass media for the highest and lowest covered aspects are (15.9%,0.8%), (9.9%,1.4%), (17%,2.3%), and (12%,0.1%), respectively. For the social media follower community, these are (26%,0.5%), (13.6%,0.6%), (14.1%,0.06%), and (31.6%,0.03%), respectively. The high inequities observed in these ordered pairs are indicative of the bias in aspect coverage exhibited by both mass media and social media.

A.2 Coding Schema

In this work, we use qualitative content analysis, along with computer based automated techniques to analyze bias in mass media, with respect to the way economic policies are represented. The use of coding schema for content analysis is evident in several works [195, 14]. On similar lines, we build a coding schema to help the annotators map policy aspects to the five constituencies of *poor, middle class, informal sector, corporate*, and *government*. The coding schema acts as a guide that helps annotators understand with minimum subjectivity, the constituencies to which an aspect should be mapped. This is done by studying the articles in that aspect manually, and then checking the schema to understand the kind of keywords, similar examples, which are relevant to the constituency. An aspect may be mapped to multiple constituencies based on the schema.

We have built this schema after careful analysis of 100 articles randomly selected for each policy event by the annotators. A total of eight annotators built this schema (for all four policies) after multiple rounds of due deliberation and cross checking. Three annotators (who were not involved in the schema development phase) were then provided

training, who went on to perform the final aspect to constituency mapping. It must be noted that out of the 11 annotators used in this process, only three belong to the author group of this paper. Thus, we ensure minimal subjectivity in the schema development and mapping process. We show a snapshot of the coding schema for Demonetization as an example. The coding schemes for all of the other policy events are created similarly. The goal of the schema is to determine whether articles of some aspect talk about a particular constituency or not. For example, in the Indian context, articles that refer to labourers who do manual work with daily wages in the construction sector are considered to be talking about the *poor* constituency, and articles that talk about workers in the IT industry, are considered to be talking about the *middle class*. To ensure consistency across different events, these definitions are suitably customized for each event. Thus, for farmer protests, articles referring to smallholder farmers and tribals are considered as discussing the poor, while articles discussing urban consumers and big farmers are considered to be talking about the middle class. This helps the annotator understand the general definition of the constituency based on that policy event.

A.3 Aspect to Constituency Alignment Matrices

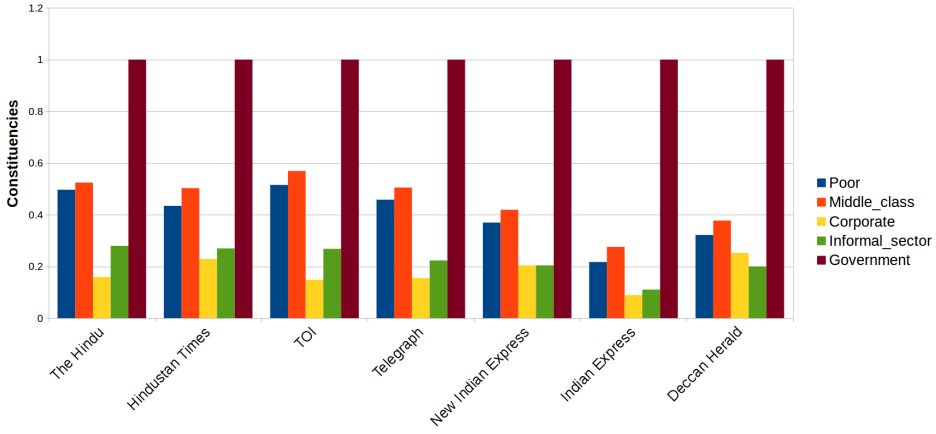
The final aspect to constituency mappings (alignment matrices U as described in chapter 5) obtained after the annotation are shown in tables A.6 to A.9. These matrices are formed after consulting the coding schema by the annotators. The value (+1/-1) in each cell indicates the alignment value for that (aspect,constituency) combination. An alignment value of 0 indicates that the aspect is unrelated to the constituency. These matrices aid us in obtaining the final stance (pro/anti) of the news-source towards the constituency as shown in equation 5.3.

A.4 Coverage of Constituencies by Mass Media

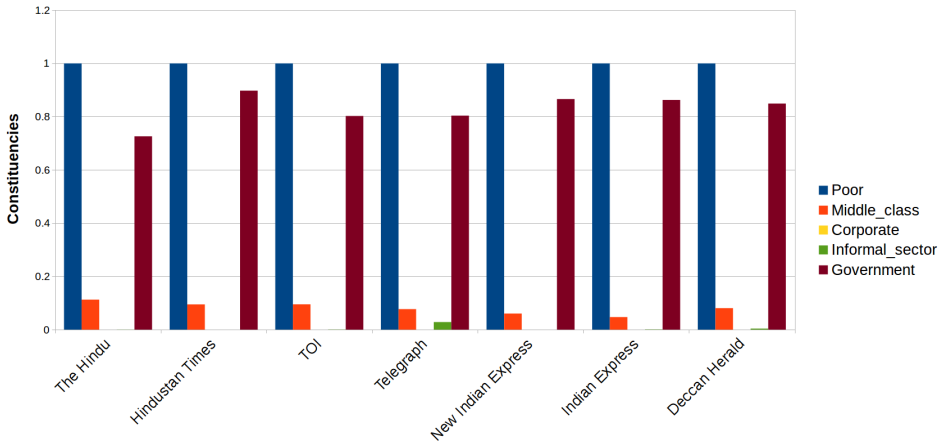
In this section, we analyze the relative coverage provided by the mass media houses to the five constituencies of *poor, middle class, corporate, informal sector, and government*. For each constituency, we aggregate the relative coverage provided to the aspects belonging to that constituency using the following equation:

$$relative\_constituency\_coverage = \frac{\sum_{a \epsilon const} w_a}{\sum_{asp \epsilon A} w_{asp}} \tag{A.1}$$

where a and asp are aspects, A is the set of all aspects for a policy event, and w_a is the total number of words across all articles for aspect a. We show the constituency coverage for Demonetization and Farmers' Protests in figure A.2. The findings for GST and Aadhaar are similar to that of Demonetization. As also stated in chapter 5, we observe that for Demonetization, Aadhaar, and GST, the coverage provided to the immediate problems of the poor are significantly lesser than that provided to the politics around a policy issue (represented by the *Government* constituency). Only in case of Farmers' Protests is the coverage provided to *poor* high. This is because most discussions on issues of farmers involve poor farmers and daily wagers, and both the ruling party and the opposition make a significant number of statements in the mass media on this sensitive issue. The least discussion happens for the constituency *informal sector*, which includes the majority workforce in India [135], and includes workers, labourers, vendors, and small traders belonging to the unorganized sector with often low levels of income. In tables A.1, A.2, A.3, and A.4 we report the KS-statistics (2-sample test) of relative coverage for each pair of the five constituencies. The high values of KS-statistics along with the low p-values indicate that the difference in coverage are significant with above 99% significance level (that is, we can safely reject the null hypothesis that the coverage come from the same distribution for two different constituencies). This also indicates the existence of a constituency bias in the mass media.



(a) Demonetization



(b) Farmers' Protest

Figure A.2: Relative coverage provided by the mass media to each of the five constituencies for Demonetization and Farmers' Protests

	Poor	**Middle Class**	**Corporate**	**Informal Sector**	**Govt.**
Poor	0	0.43	0.86	0.86	1
Middle Class	0.43	0	1	0.86	1
Corporate	0.86	1	0	0.43	1
Informal Sec.	0.86	0.86	0.43	0	1
Government	1	1	1	1	0

Table A.1: KS statistics (2-sample test) for relative coverage provided by the mass media to the five constituencies for Demonetization. All p-values lie below 0.05.

	Poor	**Middle Class**	**Corporate**	**Informal Sector**	**Govt.**
Poor	0	1	0.71	1	1
Middle Class	1	0	1	1	1
Corporate	0.71	1	0	1	1
Informal Sec.	1	1	1	0	1
Government	1	1	1	1	0

Table A.2: KS statistics (2-sample test) for relative coverage provided by the mass media to the five constituencies for Aadhaar. All p-values lie below 0.05.

	Poor	**Middle Class**	**Corporate**	**Informal Sector**	**Govt.**
Poor	0	1	1	0.57	1
Middle Class	1	0	0.43	1	1
Corporate	1	0.43	0	1	1
Informal Sector	0.57	1	1	0	1
Government	1	1	1	1	0

Table A.3: KS statistics (2-sample test) for relative coverage provided by the mass media to the five constituencies for GST. All p-values lie below 0.05.

	Poor	**Middle Class**	**Corporate**	**Informal Sector**	**Govt.**
Poor	0	1	1	1	1
Middle Class	1	0	1	1	1
Corporate	1	1	0	0.71	1
Informal Sector	1	1	0.71	0	1
Government	1	1	1	1	0

Table A.4: KS statistics (2-sample test) for relative coverage provided by the mass media to the five constituencies for Farmers' Protests. All p-values lie below 0.05.

A.5 Alignment of news-sources with their readers

In this section, we show our results for the four policies, corresponding to the research question: *Are some news-sources more closely aligned with their readers (on social media) than others?*. In figures A.3, A.4, A.5, and A.6, we report the news-source wise cumulative distribution functions (CDFs) for the sentiment distribution of mass media and social media, for all news-sources except *The Hindu*, which is present in our main paper in chapter 5. Our findings tally with the results presented in chapter 5.

Const.	**Does the article primarily target:**	**Examples**	**Normative Definition**
Poor	Labourers, workers in factories and small mills (e.g., textile and diamond-cutting mills), migrant workers and labourers, poor people belonging to the lowest level of income, and workers without bank accounts	The city at present has 9000 manufacturing units which employ 7 lakh workers. In November, by the time the demonetization was announced, at least some wages were paid and some were managed during the last two weeks ...	Poor people at the lowest levels of income. This includes labourers and factory workers without bank accounts. The welfare schemes which target the poor directly, like PMGKDS also come in the ambit of this class. Casual workers (working on contractual basis) with daily wage below 200 INR.
Middle class	Workers employed in sectors with higher income range (e.g., daily wagers working in garment based activities like stitching), ATMs, cash withdrawal limit, Note exchange at post offices, banks, customers, Long queues, city residents and office workers belonging to the service sector (e.g., working at public offices, MNCs, etc.)	The currency stock at banks across the city has improved but most ATMs still display the message temporarily unable to dispense cash. A large number of people who thronged ATMs to take advantage of the increased withdrawal limit were left disappointed on Tuesday as majority of the machines were empty.	Middle class people who suffered the immediate aftermath of the policy move like standing in long queues at ATMs, lack of money exchange at banks and post offices, and so on. Passengers of public transport. Regular workers/daily wage earners (employed on a permanent basis) with daily wage above 200 INR.

Const.	Does the article primarily target:	Examples	Normative Definition
Corporate	Manufacturing companies, industries, MSMEs, factories, multinationals, businesses, big real estate companies, entrepreneurs, businessmen, bizmen, import, export, raw material, brands, marketing, sensex, investors, NSE, NIFTY, BSE, foreign capital,	The ban on currency notes has brought business in the industrial city almost to a halt. Due to cash crunch , it has become difficult for industrialists to buy raw material, pay bills and make payments to labourers.	Big business houses, industrialists, SMEs and MNCs, and corporate business houses in general.
Informal sector and small traders	Unorganized sector, informal sector, companies not registered, unregistered enterprises, Small vendors/businesses, garment sellers, paanwallahs, chaiwallahs, small shopkeepers	The impact of demonetisation on ancillary units in the unorganised sector is likely to have a cascading impact on registered manufacturing units in the long run ...	Unorganized sector, unregistered companies, small traders, and vendors.
Government	State/Central government, state, centre, names of prominent politicians, MP/MLA, their relatives, names/positions of important government officials and designations, welfare schemes (Jan Dhan accounts)	Additionally, after a huge surge in deposits, Jan Dhan accounts witnessed net withdrawal of Rs 3,285 crore in the last fortnight ...	State and central government, policy makers, ministers, ministries, MPs, MLAs, and their relatives. Discussions in Parliament or assemblies about the narrative on Demonetization also come in this class.

Table A.5: Snapshot of the coding schema for Demonetization

Aspect	poor	Mid. Class	Corp.	Informal Sector	Govt.
Failure of RBI to stabilize economy and answer questions raised post demonetisation	0	-1	-1	0	1
Negative impact of Demonetisation on small and medium scale industries and its employees	-1	-1	-1	-1	1
Long queues at banks and ATMs and cash crunch	0	1	1	1	0
Court verdicts related to demonetisation and penalties issued for black money hoarders	1	1	-1	0	1
Vendors going cashless and negative impact on them due to demonetisation	0	-1	0	-1	1
Probes and arrests of black money hoarders	0	-1	1	0	-1
Suffering of farmers and relief measures for them	-1	-1	0	-1	1
Huge deposits in PMY bank accounts post demonetisation and announcement of minimum balance requirements	-1	-1	0	0	1
Negative impact of demonetisation on rural economy,national economy,industries,GDP,job creation etc.	0	-1	-1	0	1
Appreciation by PM for supporting Demonitisation	-1	-1	-1	-1	-1

Table A.6: Alignment matrix for Demonetization

Aspect	poor	Mid. Class	Corp.	Informal Sector	Govt.
Positive effect of climatic conditions on agriculture yield	-1	0	0	0	0
Opposition's protests and concerns on problems related to farmers (including Demonetization)	1	0	0	0	-1
Educational Drives and Social Awareness on Farmer's Distress	-1	0	0	0	0
Politics in Maharashtra/Punjab over Compensation and Loan waiver for farmers	-1	0	0	0	1
Variation in Crop Prices with monsoon season	-1	-1	0	-1	0
Cultural events related to farmers' distress	1	0	0	0	0
Disbursement of Loans and Subsidy by Banks for farmers	-1	0	0	0	1
Crimes and Suicide in farmer community	-1	0	0	0	1
Protests by farmers	-1	0	0	0	0
Loan waiver implementation by State Govts (and opposition's protests regarding the same)	-1	0	0	0	1
Training programmes and seminars to improve farming in States	1	0	0	0	1
District Administration and farmers Issues	-1	0	0	0	1
Farmers' Distress regarding Minimum Support Price for Crops	-1	0	0	0	1
Irrigation concerns and water pollution affecting farming	-1	0	0	0	1

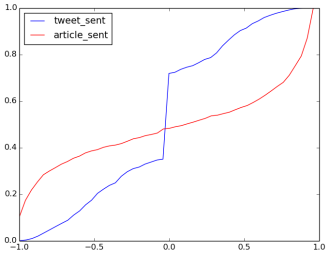
Table A.7: Alignment matrix for Farmers' Protests

Aspect	poor	Mid. Class	Corp.	Informal Sector	Govt.
Requirement of Aadhaar for passport and other services (concessions)	0	1	0	0	1
Fake ration cards caught due to Aadhaar linkage, aiding in the good of poor and middle class	1	1	0	0	1
Installation of e-pos systems for Aadhaar enabled PDS causing resentment among poor and middle class	-1	-1	0	0	1
Digitization of Aadhaar enabled employees' provident fund, attendance systems at public offices, and cashless payments helping the middle class	0	1	0	0	1
Requirement of Aadhaar for school admission and the middle class	0	1	0	0	1
Linking of Aadhaar with different schemes like PAN, mobile numbers, and bank accounts	0	1	0	0	1
Implementation of Direct Benefit Transfer Scheme	1	1	0	0	1
Aadhaar based verification for middle class telecom users and data leakage charges against telecom companies	0	1	0	0	-1
Opening of bank accounts and financial inclusion helping the poor	1	1	0	0	1
LPG Subsidy & DBT (Direct Benefit Transfer) helping the poor and middle class	1	1	0	0	1

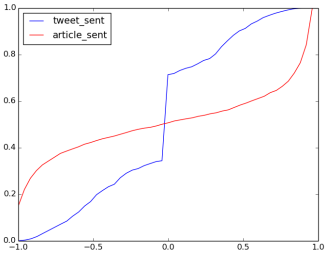
Table A.8: Alignment matrix for Aadhaar

Aspect	poor	Mid. Class	Corp.	Informal Sector	Govt.
Sensex/Market rejoices on GST roll-out	0	1	1	1	1
GST Bill implementation in State Legislative Assemblies	0	1	1	1	0
Traders hassles to meet GST registration deadline and changes in sensex/nifty	0	0	-1	-1	-1
Fears of Capital Crunch among traders before GST Rollout	0	0	1	0	1
Training Programmes for Creating GST Awareness	0	0	1	1	1
Centre-State deadlock in GST implementation	0	0	0	0	1
Uncertainity in Impact of GST on consumer Goods	-1	-1	1	0	0
Confusion amidst implementation of GST rate slabs	1	1	0	0	-1
Effect of GST on GDP and Economy	0	1	0	0	0
GST Bill Discussion in Parliament	0	0	1	0	1

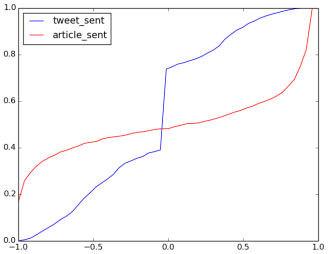
Table A.9: Alignment matrix for GST



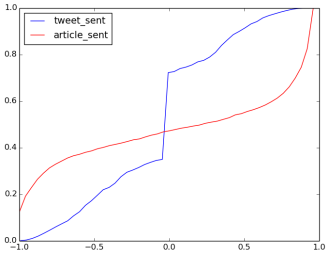
(a) Hindu



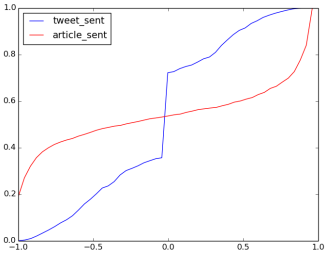
(b) TOI



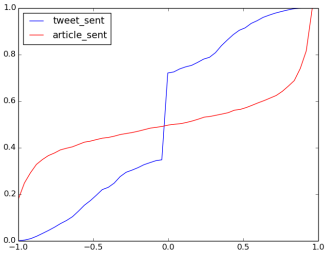
(c) TeleG



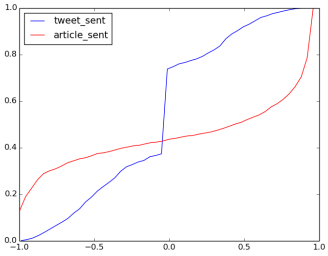
(d) NIE



(e) IE

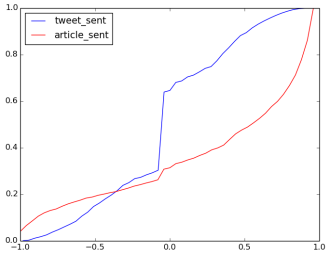


(f) HT

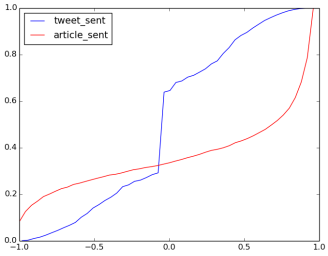


(g) DecH

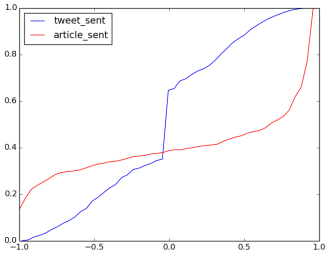
Figure A.3: CDF plots of article sentiment and tweet sentiment for the set TweetFol, for Demonetization, across news-sources.



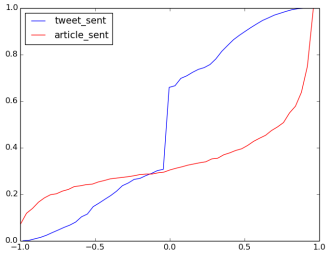
(a) Hindu



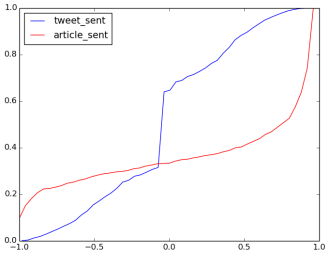
(b) TOI



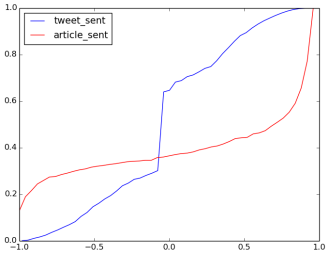
(c) TeleG



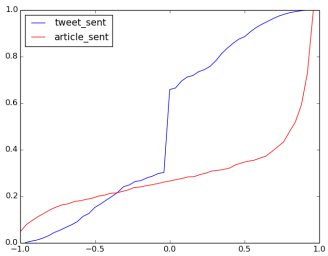
(d) NIE



(e) IE

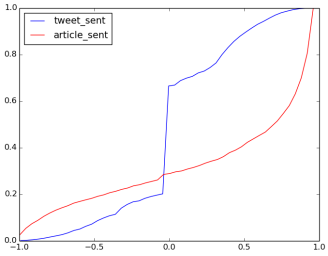


(f) HT

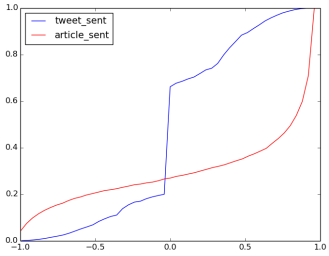


(g) DecH

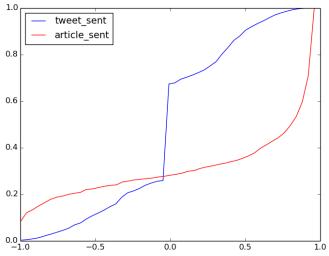
Figure A.4: CDF plots of article sentiment and tweet sentiment for the set TweetFol, for Aadhaar, across news-sources.



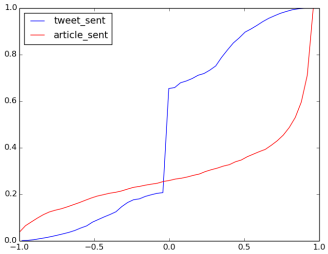
(a) Hindu



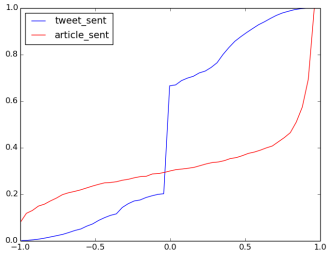
(b) TOI



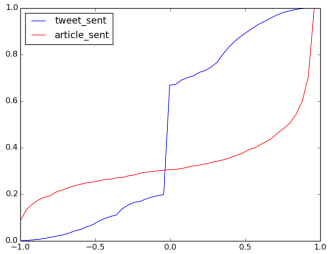
(c) TeleG



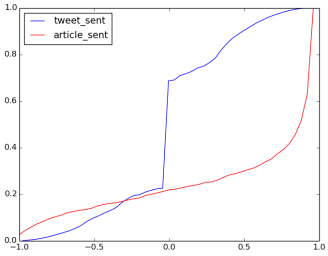
(d) NIE



(e) IE

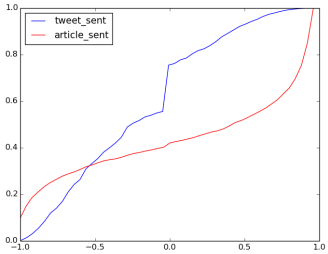


(f) HT

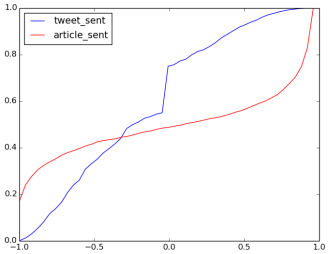


(g) DecH

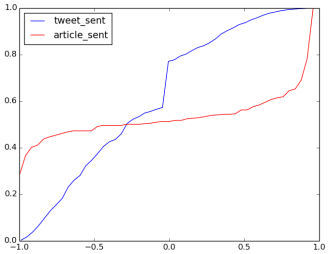
Figure A.5: CDF plots of article sentiment and tweet sentiment for the set TweetFol, for GST, across news-sources.



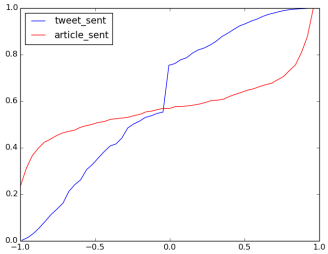
(a) Hindu



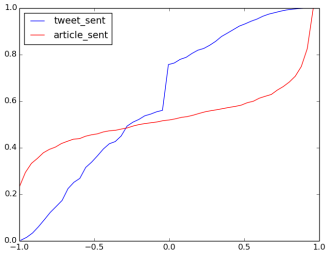
(b) TOI



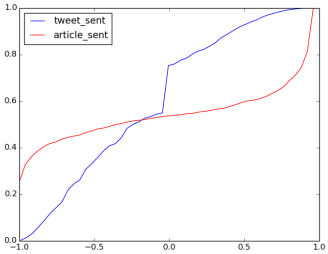
(c) TeleG



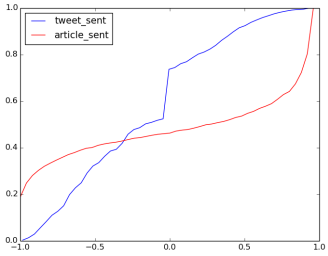
(d) NIE



(e) IE



(f) HT



(g) DecH

Figure A.6: CDF plots of article sentiment and tweet sentiment for the set TweetFol, for Farmers' Protest, across news-sources.

Appendix B

Towards a Fairness and Diversity Guaranteeing News Aggregator

B.1 Calculating the Positive Percentage

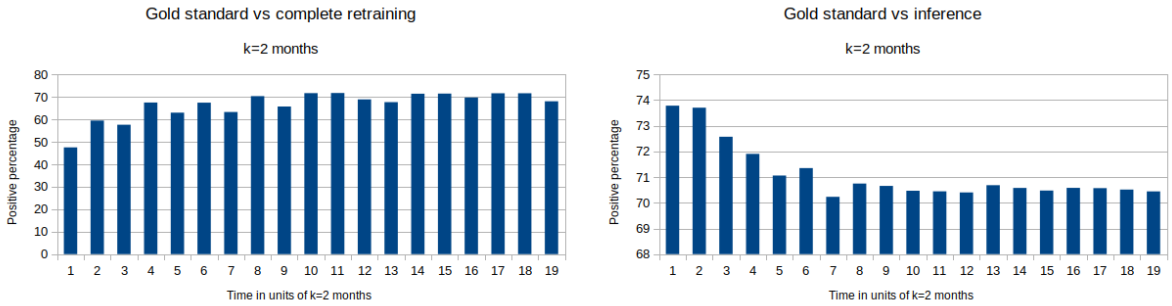
Here, we describe our technique of comparing the two LDA models, namely inferencing and retraining as described in chapter 8. There exist various methods to compare two LDA based topic models. One can use the mdiff functionality provided by the Gensim library to calculate the correlation between two topic clusters. However, this method requires setting appropriate thresholds to measure cluster similarities among two topic models. We used a simpler method to compare two topic models. Broadly, we say two topic models are similar, if they place document pairs similarly. For example, let us assume two models L1 and L2, and three documents d1,d2, and d3. The documents in our case refer to the news articles, while the topic models generate the set of aspects. The possible document pairs that can be formed are (d1,d2), (d2,d3), and (d1,d3). We now check how these document pairs are placed in L1. Let us assume that L1 places (d1,d2) in the same cluster, keeping d3 separate. In L2, let d1, d2, and d3 are all kept as separate clusters. We define positive percentage as the fraction of cluster pairs that

are placed similarly in both topic models. For this example, we see that the pair (d1,d2) are placed differently in the two models, since they are clustered together in L1 but not in L2. On the other hand, (d1,d3) and (d2,d3) are placed similarly in both models, i.e., they decoupled in both L1 and L2. Thus, we get a positive percentage of 2/3=0.67 as the similarity measure between L1 and L2. The advantage of this method is that it considers that decoupling of documents also contributes to model similarity. While obtaining topic clusters is the prime purpose of LDA based methods, the choice of keeping documents in different clusters also tells us how similar two models are in terms of their clusters.

B.2 Comparison of Inferencing and Retraining

As discussed earlier, inferencing performs much faster compared to retraining as the corpus increases in size, while retraining is more accurate as an aspect identification method. We however want to understand the extent to which retraining outperforms inferencing in terms of accuracy, for our dataset. Figure B.1 shows the plots of positive percentage for inferencing and retraining calculated using the gold model as benchmark. For these plots, we use the time interval of k=2 months for inferencing or retraining.

These plots show that initially, inferencing performs poorly as compared to the gold model, and improves slowly over time. It finally settles to a positive percentage of around 70%. On the other hand, retraining the model periodically performs much better than inferencing initially, but its performance deteriorates gradually, and settles to a final positive pecentage of 70% similar to inferencing. Thus, we see that while the performance of inferencing and retraining are different in the beginning, they tend to converge over time for our policy data. We varied k from two weeks to six months, and found similar results for all of these k-values, for all of the four policies. Hence, we stick to inferencing currently as our aspect identification approach.



(a) Inferencing for Aadhaar

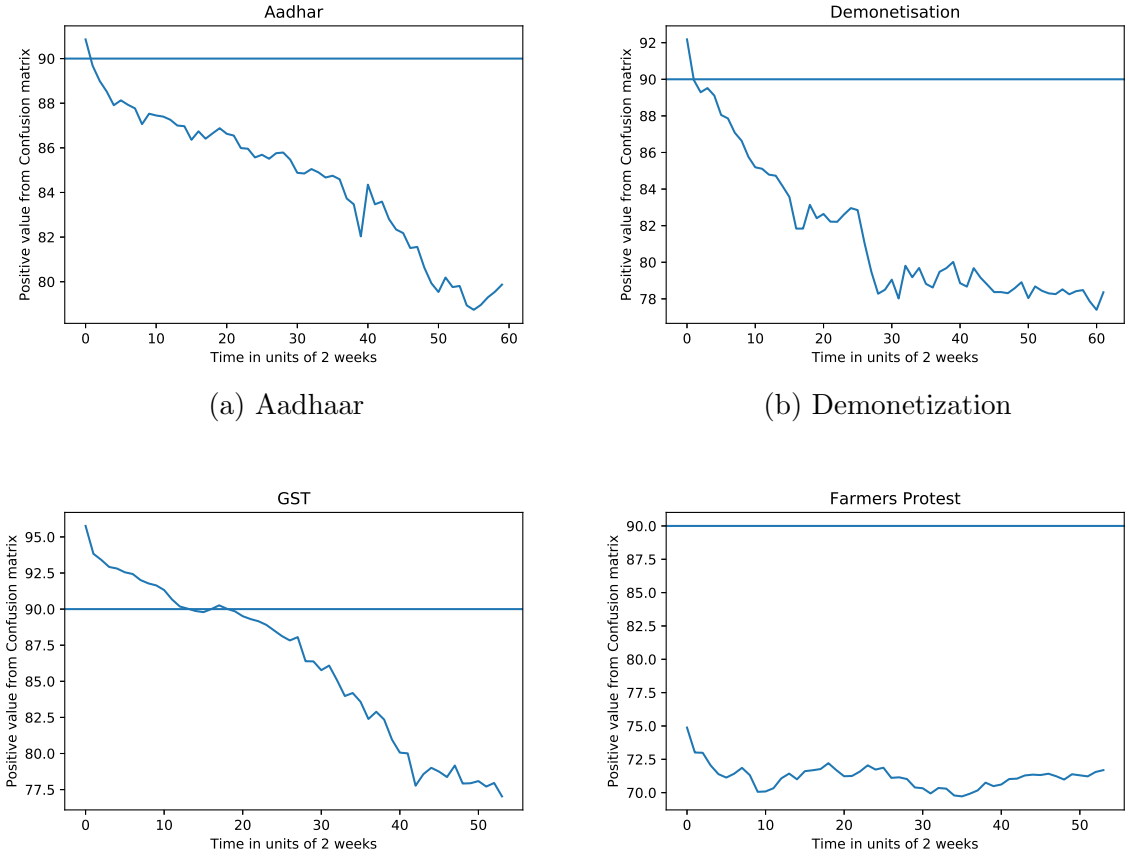
(b) Retraining for Aadhaar

Figure B.1: Comparison of retraining and inferencing schemes with respect to the gold model considering k=2 months: The positive percentage settles at around 70% for both of the schemes

B.3 Calculation of the Fairness Window

The fairness window is defined as the period of time that the aspects of a policy event take to evolve. In the beginning, an event might have a few aspects of discussion. Over time, these aspects might change – new aspects might get added to the event, while old aspects might become obsolete or merge together to form a new aspect. The fairness window thus depends on the event dataset under consideration.

To obtain an estimate of the fairness window for our four policies, we compare the performance of the two approaches of aspect identification (inferencing and retraining) directly. The idea behind this experiment is as follows: as the aspects of an event evolve, inferencing will start deteriorating in its performance compared to retraining, since inferencing works with a fixed set of aspects, and does not consider aspect evolution. Thus, if we capture the period after which inferencing starts underperforming compared to retraining, we can get an idea of the fairness window. For this purpose, we directly compare the inferencing and retraining based models based on the positive percentage, similar to our experiment in the Results section. We present our plots in figure B.2. We consider



(a) Aadhaar

(b) Demonetization

(c) GST

(d) Farmers' Protests

Figure B.2: Direct comparison of retraining and inferencing schemes: we consider a threshold positive percentage of 88% to define the fairness window, since this gives us an acceptable time period after which we can retrain the model. The minimum period for which a positive percentage>88% is maintained across Aadhaar, Demonetization, and GST turns out to be three months.

a threshold of a positive percentage of 88% for the performance comparison, i.e., if the positive percentage between the two models drops below 88%, we consider the performance gap between the two significant. Thus, the period of time for which the positive percentage between the two models (starting from the date of creation of the topic mod-

els) remains above 88% can be considered as the fairness window. From the plots, we see that this period is around three months (12 weeks) for Aadhaar and Demonetization, and 26 weeks for GST. For Farmers' Protests, the performance of inferencing is below-par throughout the timeline, and we consider it as an outlier since Farmers' Protests covers a wide spectrum of agriculture related events. Considering the other three events, we thus see that three months can be a safe choice for the fairness window. Once we consider aspect evolution in our recommendation framework, we can retrain the topic model every three months based on these findings.

B.4 Performance based on a Modification of the Algorithm

As discussed earlier, we ensure recency in our recommendation algorithm by selecting latest aspects for exposure from the last 15 days from the current feed date. However, it may so happen that all aspects generated during this period violate the upper bound constraint. In the proposed algorithm, in such cases we select the aspect (generated within the last 15 days) that least violates the constraint, and select it for exposure. In this section, we experiment with a variation of the algorithm where instead of restricting the aspect selection to the last 15 days and violating the constraint, we select older generated aspects that have not yet violated the upper bound constraint. That is, when we do not find a suitable aspect in the last 15 days, we keep checking older aspects/articles till we find one that is yet to violate the upper bound constraint. We present this modification of the algorithm in Algorithm 2. Note that the only change in the algorithm is in line 8 where we force it to select an earlier article if its aspect does not violate the constraint as discussed. We experimented with this version of the algorithm as well, and found that it still ensures fairness and diversity across all events for the parameter combination of $(f* = 0.5, d* = 0.7)$, and outperforms the baselines. However, in this case, since we put less emphasis on selecting the latest articles, the recency suffers even further as can be seen from figure B.3.

Algorithm 2: Pseudo-code to ensure fairness, diversity, and recency across news-feeds

Data: Aspects A identified using LDA within the first six months of news data available for an event

foreach $new\_day \in event\_timeline$ **do**
  $feed\_size$ = Determine the size of feed from eq. 8.3;
  $articles$ = Choose all articles belonging to this event published in the $diversity\_window$ (last 15 days), sorted in a reverse chronological order;
  foreach $a_j \in A$ **do**
    Calculate constraint U_j for aspect a_j using eq. 8.9;
  Initialise $exposure_j = 0 \; \forall a_j \in A$;
  for $k \leftarrow 1$ *to* $feed\_size$ **do**
    $item$ = Pick the latest unpicked article from $articles$ if it belongs to an aspect a_j that does not violate the constraint U_j by checking $exposure_j$. In case no such article exists, pick the latest article (generated earlier than 15 days) that does not violate the constraint. In case there still exists no such article, keep the slot empty. ;
    Place $item$ in the k^{th} position in the feed and output the feed;
    $exposure_j = exposure_j + 1$;

We see that with this version of the algorithm, the recency of our recommendation policy performs the worst. This experiment also corroborates the fact that according to our algorithm, there exists a trade-off between fairness/diversity and recency – if we attempt to enforce fairness/diversity strongly (by relaxing the 15 day window requirement of article selection, and selecting aspects that do not violate the upper bound constraint), recency suffers. On the other hand, ensuring just recency would make the algorithm unfair (as can be seen from the performance of the *Latest news first* policy. Hence, we choose to stick with the current version of our algorithm (Algorithm 1) that ensures fairness/diversity, while maintaining recency to some extent. As discussed, we plan to take up the task of improving recency further in future.

(a) Aadhaar (b) Farmers' Protest

Figure B.3: Relaxing the 15-day criteria for selecting articles in an attempt to pick under-represented aspects as suggested by U_j scores tends to significantly worsen the average feed age.

B.5 Filtering Event based Articles from Google Alerts

In this section, we describe the data collection methodology used to fetch articles from Google Alerts. First, we create a new Google account so as to prevent any personalization bias from the alerts that we receive from Google Alerts. Then, using the keywords that we had also used to gather articles from mass media for different events, we subscribe to email based alerts. We receive periodic emails from Google with news articles related to the keywords that we had specified. We get around 15208 news articles on a timeline from Feb'19 till Aug'19. These articles are then classified into either of the four events or irrelevant. We use a simple term frequency based classifier with a threshold, so that we finally have a set of 8191 articles for the four events after weeding out irrelevant articles (1068 for Aadhaar, 2566 for Demonetisation, 1392 for GST, and 3165 for Farmers' Protest). To validate the classification heuristic we randomly draw 100 articles from the set of all articles that we had received originally and manually check if the labelling is indeed correct. We find that our classification algorithm performs with 85% accuracy.

B.6 Performance on Highly Skewed Dataset

As discussed in the earlier sections, the fairness policy used by our recommendation algorithm is the minimum threshold policy coupled with production distribution of aspects. Our recommendation algorithm achieves greater fairness and diversity compared to all baselines, when considering the production distributions of the four policy events. These distributions are fairly skewed, and can act as good proxies for user specifications in future, since users tend to browse for popular news aspects.

In this section, we present our further analysis on a case study where we create a synthetic dataset, and test how our algorithm performs on it in terms of fairness and diversity with the same optimal parameter pair of (f*=0.5,d*=0.8), for highly skewed platform preferences. The production distribution P is replaced with these preferences (skewed distributions) in our minimum threshold policy for this study. We consider the two policy events of Demonetization and Aadhaar, and cluster their articles into three aspects indicative of their political ideology: pro-policy (where the articles support the policy move), anti-policy (where the articles criticize the policy move), and neutral (where the articles are objective and do not significantly criticize or support a policy event).

The reason behind using political ideology of articles to define aspects is that generally, users tend to perceive political news on their ideological frames. In a real-world scenario, it would be much easier to obtain user requirements or platform preferences based on political ideologies, since automatically identified aspects using LDA are much fine-grained and may not be understood well by all. Additionally, while LDA identifies different aspects for each event, political ideologies remain fixed across events if the policies are implemented by the same government. Thus, using the ideological classification (pro vs. anti establishment), we can use our news recommendation framework combining all policy events, instead of the current approach of addressing one event at a time. We classify a subset of 500 articles from the two policies manually with the help of four annotators. The inter-coder agreements (using cohen's kappa) of the article mapping exercise were 75% for Aadhaar, and 79% for Demonetization. The high values of agreements are in-

dicative of the reliability of the mapping exercise. As part of our future work, we are also developing an ideology classifier, which can classify the entire media corpus automatically to these three ideologies.

After the annotation exercise, we obtain 350 pro-policy articles (250 for Aadhaar, 100 for Demonetization) and 50 anti-policy (35 for Aadhaar, and 15 for Demonetization) articles. We next apply our news recommendation algorithm on the unified set of 400 articles for Demonetization and Aadhaar, and produce feeds by varying the distributions on two ideologies: pro-policy and anti-policy. We remove the neutral articles from this set, in order to keep our study simple. Note that unlike the original study where we consider news articles actually produced on a daily basis, here we spread the pro and anti articles uniformly over a simulated timeline of four months, since there might be dates when no pro/anti policy articles are produced. In other words, we consider a production of four articles per day, and distribute the articles evenly on the imaginary timeline of 90 days, to simulate daily production of pro/anti policy articles. We start generating the feeds using our algorithm for the last 15 days, and consider various combinations (starting from 100% pro-policy requirement to 100% anti-policy requirement, with a change of 20% for each distribution) of distributions of aspect coverage. We obtain the plots shown in figure B.4 for our analysis.

From the plots, we find that our algorithm successfully ensures fairness and diversity in aspect exposure, even for these multiple skewed platform preferences, and outperforms all of the baselines for the same pair of optimal parameter values $((f* = 0.5, d* = 0.8))$. This proves that for our dataset, the recommendation algorithm can ensure fairness and diversity in most scenarios, while also giving importance to the distributions specified by platform managers.

B.7 Results (continued)

We show in figure B.5 the plots for recency for the remaining two events (in continuation

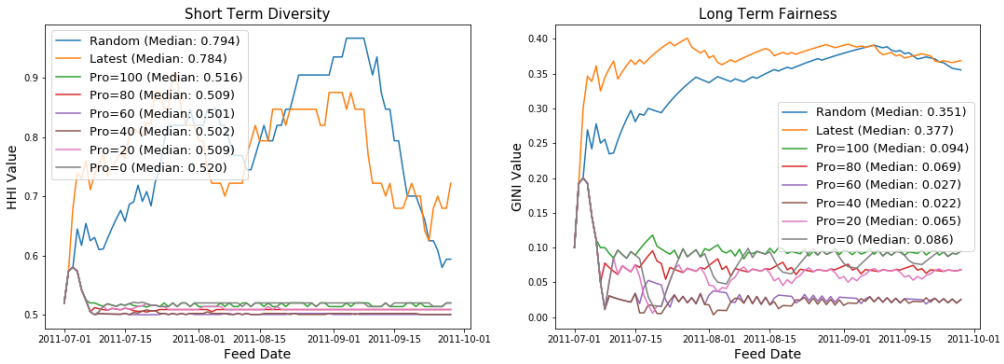


Figure B.4: GINI and HHI plots for Aadhaar+Demonetization for pro and anti policy articles. For the optimal parameters of $(f* = 0.5, d* = 0.8)$ our algorithm outperforms the two baselines in terms of fairness and diversity.

of figure 8.6), i.e., Demonetization and Farmers' Protests. Our findings are the same as explained earlier. In figure B.6, we present the results of our recommendation algorithm

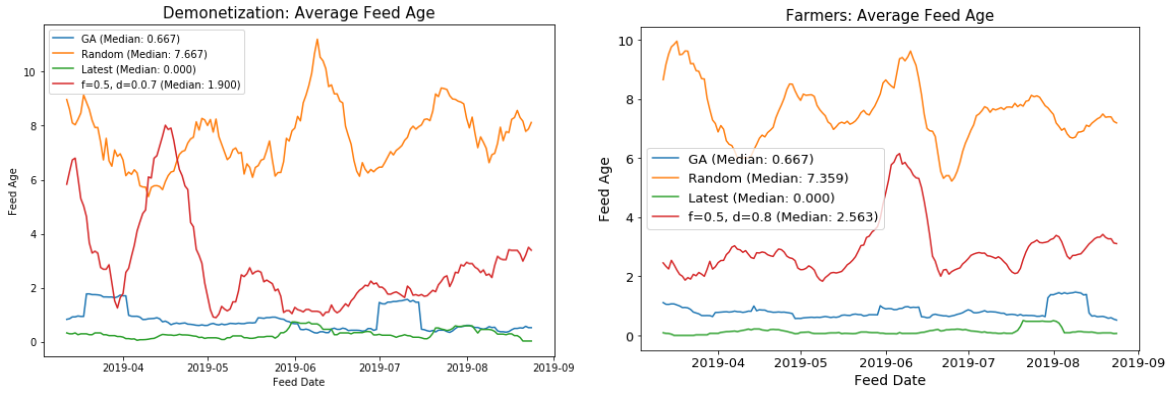


Figure B.5: Weekly average news-feed age for our recommendation algorithm and the baselines, for Demonetization and Farmers' Protests

in terms of fairness and diversity for Demonetization and Farmers' Protests (continuing from figure 8.5). We see that the findings remain the same: our algorithm outperforms all of the three baselines for the four events in terms of fairness and diversity.

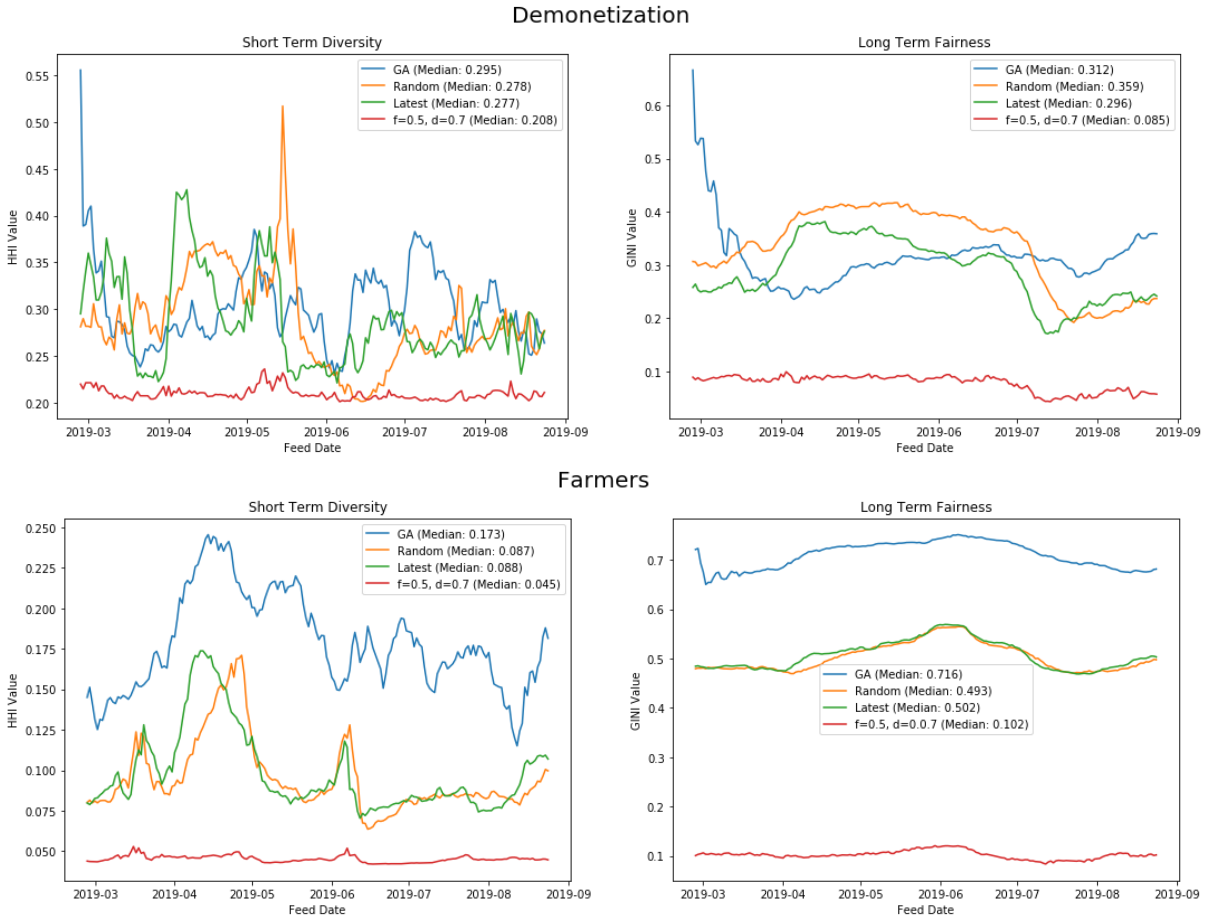


Figure B.6: GINI and HHI Plots for Demonetization and Farmers' Protest: our algorithm is seen to outperform all of the baselines for its optimal combination of parameters ($f* = 0.5, d* = 0.8$)

List of Publications

1. "What Drives Location Preference for Corporate Social Responsibility (CSR) Investments in India?" Varun Pareek, Rohit Sharma, Anirban Sen, Arundeep Gupta, Manikaran Kathuria, and Aaditeshwar Seth, ACM COMPASS 2020, Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies

2. "Studying the Discourse on Economic Policies in India Using Mass Media, Social Media, and the Parliamentary Question Hour Data", Anirban Sen, Debanjan Ghatak, Kapil Kumar, Gurjeet Khanuja, Deepak Bansal, Mehak Gupta, Kumari Rekha, Saloni Bhogale, Priyamvada Trivedi, and Aaditeshwar Seth, ACM COMPASS 2019, Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies, 2019

3. "An Attempt at Using Mass Media Data to Analyze the Political Economy Around Some Key ICTD Policies in India", Anirban Sen, Priya Kumari, Pooja Aggarwal, Manpreet Singh Agandh, Aditya Guru, Debanjan Ghatak, Sravan Verma, and Aaditeshwar Seth, ICTDX 2019, Proceedings of the Tenth International Conference on Information and Communication Technologies and Development, 2019

4. “Empirical Analysis of the Presence of Power Elite in Media”, Anirban Sen, Pooja Aggarwal, Aditya Guru, Deepak Bansal, I Mohammed, J Goyal, K Kumar, K Mittal, Manpreet Singh, M Goel, S Gupta, Varuni Madapur, Vipul Khatana, and Aaditeshwar Seth, ACM COMPASS 2018, Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, 2018

5. “Leveraging Web Data to Monitor Changes in Corporate-Government Interlocks in India”, Anirban Sen, A Agarwal, Aditya Guru, A Choudhuri, G Singh, Imran Mohammed, J Goyal, K Mittal, Manpreet Singh, Mridul Goel, S Gupta, S Pathak, Varuni Madapur, and Aaditeshwar Seth, ACM COMPASS 2018, Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, 2018

Biography

Anirban Sen is a PhD candidate at the Department of Computer Science and Engineering at the Indian Institute of Technology, Delhi. He joined the institute as a PhD student in 2014, and has a Master's degree in Computer Science and Engineering from Indian Institute of Engineering Science and Technology, Kolkata. His research interests include computational social science, big data analysis for development, and ICTD. His work specifically focuses on building information tools to monitor different aspects of the political economy around policy events, with the help of web based data.