

**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

©Indian Institute of Technology Delhi - 2021
All rights reserved.

**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 her 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.

A handwritten signature in black ink that reads "Aadit". The signature is written in a cursive style and is underlined with a single horizontal stroke.

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 कॉर्पोरेट 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 एल्गोरिथम 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 <i>The Hindu</i> . 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 <i>degpol</i> for the top 20 entities with highest coverage: the aggregate sentiment/ <i>degpol</i> is calculated as the sum total of the values corresponding to the statements made by an entity. Higher the value of <i>degpol</i> (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. <i>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.</i>	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 pre-processing 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 pre-processing of the articles was done.	110
7.2	Relative coverage in percentage for entity groups (considering both <i>about</i> and <i>by</i> 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