Media Bias Analysis across important policies of India

Debanjan Ghatak

Indian Institute of Technology, Delhi

Media Bias Analysis across different media houses

A thesis submitted in partial fulfillment
of the requirements for the degree of

MASTER OF TECHNOLOGY

in

Computer Science & Engineering

by

DEBANJAN GHATAK(2017MCS2079)

Under the guidance of

Dr. AADITESHWAR SETH

Indian Institute of Technology, Delhi

Certificate

This is to certify that the thesis titled **Media Bias Analysis across different media houses** submitted by **DEBANJAN GHATAK** for the award of **Master of Technology** in **Computer Science and Engineering** is a record of bona fide work carried out by him under my guidance and supervision.

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. The Thesis fulfills the requirements and regulations of IIT Delhi and in my opinion meets the necessary standards for submission.

DR. AADITESHWAR SETH
Department of Computer Science and Engineering
Indian Institute of Technology, Delhi

Abstract

We examine the bias of different media-houses(newspapers) towards the four economic policies of India.The four economic policies that we analyse in our thesis are **Aadhaar,Demonetisation,Goods and Service Tax(GST) and Farmer's Protest(FP)**.Our work is significant in creating a standardized methodology to assess the following:what are the coverage given by different media-houses towards different aspects(topics) of the above mentioned policies,their sentiments towards different policies,comparison across mass media,social media and parliamentary Question Hour data,differences in sentiments expressed between mass media and social media.We are able to draw some meaningful conclusions like how some media-houses differ with respect to each other in terms of their ideologies and how the ideological slants are varying with the other forms of media.Mostly we find that the social media and mass media are aligned in the coverage given as well as in their sentiments.Through more analysis we are trying to create a model to explain the coverage given by media.

We also analyse the Parliamentary Question Hour data.Mass media covers different constituencies(poor,middle-class,corporate,informal sector and government) but shows biases towards a particular constituency for different policies.On the other hand we find that the Parliamentary Question Hour data focuses mainly on the procedural aspects of the policies and not much on the issues of poor and middle-class.Social Media displays similar stance as that of Mass Media.

Further to this,we also analyse the four ICTD policies of India(**Aadhaar,Digital India,E-Governance,Cashless Economy**).We explore the political economy around them by collecting the data about these policies in Mass Media.We find out that the politicians are most vocal regarding these policies and mostly they express similar ideologies.We also find out that the mass media are more concerned about the issues of middle-class,talking mostly about the negative aspects of these policies and also about the issues faced by the poor sector of this country.Our key contribution is a methodology of

using automated analysis of mass media data to reveal the factors that might be shaping the political economy behind policy making.

Towards the end of our thesis,we analyse the opinion articles(opeds) of mass media.These are the articles written in the opinion category of every newspaper.We do a coverage,sentiment and an entity based analysis on these opinion articles as well.

Acknowledgments

This dissertation would not have been possible without the assistance,help and guidance of several people who have helped me throughout the course of our project.
I would like to express our heartiest gratitude to my project guide,Dr. Aaditeshwar Seth for being a guiding light throughout the duration of my thesis and providing us with the opportunity of working on such interesting areas and research questions.
I would also like to thank Anirban Sen for his continuous support and mentorship during the entire tenure of our thesis.I would like to thank all my senior groups for laying the foundation of such an interesting work.I would also like to thank my project partner Gurjeet Singh Khanuja, my friends Kapil Kumar and Deepak Bansal,my juniors Rekha Kumari and Mehak Gupta for their help as and when required.

DEBANJAN GHATAK(2017MCS2079)

Department of Computer Science & Engineering

Indian Institute of Technology, Delhi

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Problem Statements:

We examine the four economic policies of India.We try to analyze that how differently are the newspapers covering these policies and what are the problems and issues about these policies they focus the most on.People form an opinion reading newspapers,so it is imperative to know how do the news-sources cover these policies.There can be different forms of biases which can exist in media(what is the coverage given to different policies across different aspects of those policies(coverage bias),how are those aspects covered per policy by different media-houses(sentiment bias).We analyze all of these biases in details for seven national English newspapers(**The Hindu,Hindustan Times,Telegraph,Times Of India,Indian Express,New Indian Express and Deccan Herald**) and also try to compare the coverage bias and sentiment bias with social media and Parliamentary Question Hour data.

Different media houses have different ideological slants(**pro-poor,pro-middle class,pro-government,anti-government**).This creates biases in Mass Media since many news-sources are biased towards some political parties,political ideologies.We try to examine and analyze these biases with a proper structured framework for different policies(events).We try to analyze how are these constituencies covered across different media for different policies.In our thesis we analyze different media biases for four economic policies.We briefly describe all these policies here.**Aadhaar:**An initiative by the government to give every Indian resident a biometric-based unique identification number. The data is collected by the Unique Identification Authority of India (UIDAI), a statutory authority under the jurisdiction of the Ministry of Electronics and Information Technology. The issue has been criticized owing to lack of security and privacy in citizens data collection and storage mechanisms, and also because of an allegedly faulty implementation of the platform

or use of the platform by different agencies.**Demonetisation:**A policy event where the government on 8 November, 2016 banned all 500 INR and 1000 INR banknotes with the motive of curtailing the use of illicit and counterfeit cash used to fund illegal activity and terrorism. The move was widely criticized owing to multiple problems caused to common people due to sudden depletion of liquidity, irregularities in norms of exchanging old currency notes, cash exhaustion in ATMs, and so on.**Goods And Service Tax(GST):**An indirect tax levied in India on the sale of goods and services. It is levied at each step of the production value-chain with an effort towards formalization in the industry and simplification of multiple types of taxes which preceded the GST regime. Since its implementation there have been intense de- bates though on its complexity and problems in implementation which have impacted the overall growth of the economy.**Farmers' Protest:**A series of protests by farmers in India including the ones at Madhya Pradesh (Mandsaur protest) and Maharashtra (Kisan long march) demanding better prices for production of crops, loan waivers, and forest rights, among others. The issue is highly active politically with a lot of involvement of different politicians and political parties.For each of these policy events, we extract a set of aspects (commonly discussed topics) from the articles published by news-sources using Latent Dirichlet Allocation (LDA). We try to analyze the following research questions in our thesis.

- Are news-sources biased on the amount of coverage they give to different aspects about the policy issues?
- Are some news-sources more closely aligned with their readers (on social media) than others?
- Which aspects and constituencies about the policies do the mass media,social media, and QH data cover?
- What are the ideological slants of mass media,social media, and QH data for these policies?

We also analyze the four different ICTD policies.We perform this analysis because policy making is known to be influenced by several factors.**Aadhaar** is

also an ICTD policy.We describe the other three ICTD policies here.**Digital India:**is a campaign launched by the Government of India,which includes plans to connect rural areas with high-speed Internet networks.**Cashless Economy:**aims to create an economic state whereby financial transactions are not conducted with money in the form of physical banknotes or coins, but rather through the transfer of digital information between the transacting parties. Among other objectives,the demonetization move was positioned as a policy to push India towards a cashless state, so that the poor who do not have any credit history or access to banking channels, will be able to create this data trail that will help them later get easier access to formal sources of credit and other financial instruments.**E-Governance:**is an initiative of the Government of India to make all government services available to the citizens of India via electronic media, instead of them having to fill up paper forms.We try to answer the following research questions:-

- Who are the top covered entities(politicians,academicians,activists,etc) in mass media?
- What is the ideological slant of these top covered entities in mass media?

We also analyze the opinion articles of mass media for different events and try to find the coverage bias related to different aspects.We also analyze the coverage and sentiment bias of the top covered entities along with analyzing the sentiment bias of the top covered authors writing those opinion articles.

1.2 Thesis Outline:

The remainder of our thesis is organised as follows.Chapter 2 contains the Related Work.Chapter 3 contains the system design, architecture and algorithms.We also explain our media analysis database and our website in this Chapter. Chapter 4 and 5 contains our analysis on the economic policies.Chapter 6 contains our analysis on the ICTD policies.Chapter 7 contains our analysis on the opinion articles.Chapter 8 contains the conclusion and the future discourse of work in this project.

Chapter 2

Related Work

We have analysed media bias in following parts:a)Mass Media Bias b) Social Media Bias c) Bias in the questions asked in Parliament d) Analysis of Political Economy around policies.

2.1 Mass Media Bias:

Journalists and news-sources shape public opinion by intentionally creating bias in their selection, writing, and distribution of news content, and for this reason they have often been called gatekeepers [1]. Chiang et al. [2] bring out evidence of endorsements provided to political candidates by mass media in the USA. [3] similarly developed an index to define a measure of media slant by analyzing key phrases in news content specific to political ideologies. Munson et al. [4] similarly in their work, assign a political bias score to each media outlet based on whether liberal or conservative candidates are over or under represented in these outlets. Budak et al. [5] 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 ideological slant of news-sources, their political affiliations in terms of the political entities they cover, and their alignment with their audience on social media.Mullainathan et al. [6] suggest such a model and posit that competition in media results in lower prices but a bias towards readers preferences. They observe that reader heterogeneity is more important to address media bias, than competition among news-sources. Prat et al. [7] suggest a similar model in their work, and go on to show that political accountability and voting outcomes are both affected by bias in mass media. These models have however not been rigorously evaluated on actual data. In our current work, we do not

delve into models to predict mass media bias, but our contribution towards creating a standard frame- work can help in the evaluation of such models which we intend to pursue as part of future work.A study by Stromberg et al. [8] observes how content targeted at selective groups of audience to maximize profitability by catering to advertiser interests affects public policies, because it amplifies the outreach of politicians to these audience groups. In the study [9], Bartels finds a significant influence of mass media exposure in opinion shift of the public in the 1980 US presidential elections. Contrarily, the study by Lazarsfeld [10] shows minimal influence of mass media on public opinion in terms of their voting choices.Our work helps understand possible biases in the mass media itself, but we also use data from analysis of the mass media to understand different factors that might be influencing the policy making process.Scheufele et al. discuss the concepts of agenda setting, framing, and priming in mass media [11], which together play a significant role in influencing public opinion on socio-political issues.Oliver and Myers [12] write that the claim of the media being an objective and neutral communicator of events has been rejected by scholars of the media for quite some time now (e.g., Herman and Chomsky [13]).Significant research has been conducted in the area of bias detection in terms of the political ideology of news-sources, blogs,and political speeches using partisan tokens [14], bag of words [15],HMM based models [16], hierarchical topic modeling [17], and deep neural networks [18]. Unlike these studies, our work focuses on probing variation in coverage of issues, and the way it engages in the policy process by way of covering different aspects relevant to the policy.

2.2 Social Media Bias:

Social media platforms have been argued to create biases in content distribution and display.Online 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 and which are amplified further algorithmically. Eli Pariser [19] discusses the issue of personalisation on the Internet through Googles segregation of its user base into

different filter bubbles. Based on the search history and many other features extracted for each user, Google produces a customized content set for each user separately, inherently biasing information presentation. Bias is also reflected in the web browsing history of users as studied in the US scenario by Flaxman et al. [20] - the authors find that most people read descriptive reporting and visit a handful of ideologically similar news outlets, leading to less exposure to content from diverse political alignments. Similar biases have been noticed in the linking pattern of social media such as blogs [21] which leads to echo chambers. Such a view is corroborated by [22] 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 [23] and also noticed in [[24], [25]]. In the absence of algorithms to support such diverse information sharing however,users need to diversify their own networks to be able to get a wider perspective as noticed on Twitter from the followership network of journalists and different media sources [26]. In our current work, we do not evaluate whether echo chambers exist in the Indian social media, or whether online social network platforms accentuate bias through algorithmic filters - we only compare the coverage and sentiment expressed on social media with that on mass media to contrast the two spheres of expression of public opinion.As the pressure towards democratization increases, it is also expected that political leaders will increasingly turn towards using social media [[27],[28],[29],[30]].An example of this was witnessed during Narendra Modis campaign in the 2014 Lok Sabha election.Pal et al. show how Modi used social media to connect with younger voters and rebrand himself from a controversial personality to a tech savvy politician in tune with the aspirations of his constituents [[31],[32]].Given that social media can play an important role in influencing and conveying public opinion, and that a majority of the population on social media is educated, we want to find out if indeed the information seen on social media appears to be more representative of citizen issues, and is different from mass medias.

2.3 Bias in the questions asked in Parliament:

While the issue of parliamentary accountability is ridden with complexities, one way to contextualize the proceedings of the Question Hour (QH) in the Parliament is to understand the policy focus and interests of sitting MPs. The website of the Lok Sabha lays down the purpose of parliamentary questions as follows: *Through the Question Hour the Government is able to quickly feel the pulse of the nation and adapt its policies and actions accordingly. It is through questions in the Parliament that the Government remains in touch with the people in as much as members are enabled thereby to ventilate the grievances of the public in matters concerning the administration. Questions enable Ministries to gauge the popular reaction to their policy and administration. Questions bring to the notice of the Ministers many loopholes which otherwise would have gone unnoticed.Sometimes questions may lead to the appointment of a Commission, a Court of Enquiry or even Legislation when matters raised by Members are grave enough to agitate the public mind and are of wide public importance.* There are a number of studies regarding parliamentary questions in democracies, which try to understand the incentives of asking a large number of questions on certain topics - whether they are linked to gender [33],caste [34],direct electoral links [35],experience or seniority [36],or constituency demographics [37].Thus, parliamentary questions can be effectively studied by classifying data with respect to the area of concern being represented. The policy focus of questions form a proxy for understanding the actual priorities and concerns of MPs in the House. While the extant literature on parliamentary questions in India is focused on quantifying involvement of individual MPs, the analysis of the question content remains largely unexplored. Our work fits in this space by looking at the questions asked regarding different policies and aspects under these policies, to see if it sufficiently represents the concerns of all sections of people.

2.4 Analysis of Political Economy around policies:

Literature on policy making has proposed frameworks to design and conduct analysis on the political economy around policies. Moncrieffe and Luttrell [38] propose a framework for such analysis for bi-lateral donor organizations. This framework describes how interactions between different organizations and their members shape policy making and outcomes, and pointers on how staff can use the analysis to assess appropriateness of policy interventions, their content and timing, and new spaces for enagagement.Holland et al. [39] propose a framework and tools for institutional, political, and social analysis of policy reforms at macro and micro levels. This web based tool aids users to analyze the rules people develop in political, social, and economic spheres; the structure of power relations that aids policy making; and the social relations that govern interactions between households, communities, and social groups. These theoretical frameworks often introduce the user to the main stakeholders involved in the policy formulation. Our work is related and lists out the key stakeholders and their interconnections.We, however, do it in an automated manner using different computer science techniques of natural language processing, to be able to apply the framework to analyze potentially any policy event.Hai et al. [40] study the process of policy formulation in developing countries, and identify a set of relevant actors who contribute towards it. This set includes elected officials, appointed officials, business actors, labour, public, think-tanks and research organizations, political parties, mass media, and interest groups. Devesh Kapur [41] in his work discusses about how policies are shaped in order to favor specific caste and class groups in India. A related work of ours in this area [[42], [43]] takes an automated approach to produce social network data of influential corporate and government entities that can be analyzed to understand the political economy behind the policy making process.

Chapter 3

System Design

3.1 Data Collection:

3.1.1 Crawlers:

To find information on the hundreds of millions of Web pages that exist, a search engine employs special software robots, called **spiders**, to build lists of the words found on Web sites. When a spider is building its lists, the process is called **Web crawling**. In order to build and maintain a useful list of words, a search engine's spiders have to look at a lot of pages.

Crawler identifies all of the hyperlinks on the website and adds them to the list of URLs to visit (also known as the **crawl frontier**). Once a web crawler is given a list of URLs (also called **seeds**) to check out, it begins visiting each website and downloading the content.

We have built a corpus of media articles published by Indian English-language news sources by crawling their web archives / RSS feeds. This system can be used for analyzing media articles, getting relevant articles by efficient querying or integrating the system with other data sources to mine interesting patterns.

We have collected news articles of categories like **national, international, regional, sports, opinion and business**. This categorization is followed by all news sources in general. National, international and regional are political news articles. Some articles are also tagged as FRONT PAGE if this information is available while crawling their urls. Sometimes URLs are published under multiple categories by news source. So we associate an array of categories for each article. Opinions include editorials, regular columns, letters to the editor and special edition columns. Currently, we have data from sources **The Hindu, The Times of India, Indian Express, The New Indian Express, Telegraph, DeccanHerald and Hindustan Times**.

Some news sources had their articles archived online which could be crawled by any user. Archives were available online for all sources present in our system except **Hindustan Times**.They are crawled into our system.We use Scrapy crawler to crawl archives.

News sources publish latest articles through RSS feeds. We have setup an automated script that uses RSSfeedparser to extract current data. It checks for new feeds every five hours. HTTP header fields like ETag and Last-Modified date are used by some servers as a property of their feeds. ETag is a unique id given to the feeds by the server. If a feed gets updated, value of ETag changes. So when we send these fields as headers, server sends a response only when there is a change in ETag or Last-Modified date of feeds. RSS feeds are available for Hindustan Times ,The Hindu ,Indian Express, New Indian Express, The Times of India only. For the rest of two newspaper i.e. Deccan herald and Telegraph, we use their web archives to fetch new data.

We store meta data of an article URL and information about a news source in a Postgres relational database. News source information like **name, language of content, country of circulation** is stored. Article information like its **title, web link, date of publishing, news source and category** is stored.

Article text is fetched separately by accessing the URLs stored in relational database.In this process, we fetch text and author information if available. Authors are reporters and columnists. This data is stored in MongoDB which is a no-sql database. An article document in MongoDB consists of its meta data which is stored in postgres along with text and author info. An article document in MongoDB has fields as shown in the table 3.1:

3.1.2 Event wise Article Extraction:

Till date we have a corpus of more than 48 lakh articles getting strengthened day by day by our live crawlers.This corpus contains articles since 2011.Once we have to start analysing an event,we form a base set of keywords selected manually for extracting event specific articles from the main article dataset

Column	Description
articleTitle	Title of the article
articleURL	Web link of the article
categories	Array of categories of an article
publishedDate	Date of publishing
publishedTime	Time of publishing
text	Text of the article
author	Information about authors who are usually reporters and columnists
sourceName	News source
country	Country of circulation of news source
language	Language of article
entities	Extracted using Open Calais
sentiments	Sentiment of the article text
keywords	Relevant topics of articles

Table 3.1: Relational database schema: Articles stored

and make a base dataset for the event.But with this process we may get many irrelevant articles for the specific event and it also may happen that we miss out on a relevant article due to missing out on a relevant event specific keyword.

So we go through the articles collected and collect a list of top keywords found using **Rapid Automatic Keyword Extraction(RAKE)** algorithm with following parameters: minimum characters in a word = 3, maximum words in a phrase = 5 and minimum frequency of a keyword = 1. Complete Algorithm is described in 1. We have defined some rules for filtering out irrelevant articles specific to an event.

- **Frequency of *keyword_set*:** If an article is relevant to an event,then we normally find the keywords for the event to be mentioned quite frequently.We replace all the keywords for the event with a common keyword and we calculate its frequency in the concerned article.It the frequency is greater than or equal to 3, then we keep the article as a relevant article.It the frequency is one or two,then we apply the following rules:
  - If we find the keyword present in the top 50 percent of the arti-

cle,then we keep the article as a relevant article.

- Else,we find out the sentences containing the keyword and we see whether those sentences are talking about the event.Stanford NLP(SNLP) is used for implementing this.

- **Keyword Extraction using RAKE**: Relevant keywords are found out using RAKE.We also extract the relevance score for all the keywords extracted.We discard all those keywords whose relevance score is less than 20% of the score of top-ranked keywords.We then write out the top 100 relevant keywords to a file and then select all the event related keywords and build an augmented keyword set.This algorithm continues until we don't find a new keyword.

- **Find bigrams that have at least one word from *keyword_set***:We find out the bigrams from the concerned article which contain atleast one keyword.We then manually find out all the unrelated bigrams and remove all the articles containing them.For instance, bigrams like GST Roads, GST Railways are found in articles related to GST event which are actually unrelated. Such articles are removed from our event specific article collection.

After this entire process,we obtain a collection in our MongoDB with all the event-specific relevant articles. We have analysed 7 different events(4 economic policies and 4 ICTD policies).The keywords for the four economic policies are listed in the table 3.2 :

The keywords for the four ICTD policies are listed in table 3.3:

3.1.3 Question Hour data:

Parliamentary Question Hour(QH) data was obtained from the website of Lok Sabha.We obtained **Lok Sabha 15 and Lok Sabha 16** data.The schema of the QH data is shown in table 3.4:

Algorithm 1 Event Article Extraction

Input : EventRelatedKeywords, MainEventWord
Output: Event_Dataset

Step 1: Keyword based article extraction
1: Find articles with text having any *EventRelatedKeywords* from our main collection.
2: Store all found articles and their information in new collection Temp_Dataset.

Step 2: Find all Bigrams that are not related to the event but have at least one word from *EventRelatedKeywords*. [e.g. GST Roads, it has word GST but is not about GST Event]
3: Initialize EventBgms , NotRelatedEventBgms as Empty
4: Tokens = nltk.word_tokenize(Temp_Dataset)
5: AllBgms = nltk.bigrams(Tokens)
6: **for each** Bgm in *AllBgms* **do**
7: **if** ntersect(Bgm.split(),*EventRelatedKeywords*) not NULL **then**
8: Add Bgm in *EventBgms*
9: **end if**
10: **end for**
11: Manually go through these EventBgms and find Bigrams that are not related to event and store them in *NotRelatedEventBgms.*

Step 3: Find articles that are in context of the topic
12: Create new collection *Event_Dataset.*
13: ConsideredRelations = [amod, nmod, dobj, iobj, nsubj, nsubjpass]

3.1.4 Tweet Extraction:

To perform analysis on social media(twitter),we need an enriched tweet dataset. We have built a system for collection of tweets.By newspaper communities we mean the followers of the official twitter handles of the 7 national newspapers mentioned before.

Most of the data collection was done using **Twitter REST API** and its python implementation **Tweepy**. To use these, authentication tokens and keys are required. The API(or Tweepy) returns the tweets in the form of tweet/status objects which can be handled and stored as **JSON ob-**

for each article in *Temp_Dataset* **do**
    for each nrbgm in *NotRelatedEventBgms* **do**
        if nrbgm in article **then**
            reject the article
            continue
        end if
    end for
    Replace all *EventRelatedKeywords* with MainEventWord.
    Find Frequency of *MainEventWord* in article as keyfreq
    if keyfreq > 2 **then**
        Store the article in *Event_Dataset*
    else if *MainEventWord* in top 50% lines of the article **then**
        Store the article in *Event_Dataset*
    else if *MainEventWord* is present in any of the *ConsideredRelations* in the article **then**
        Store the article in Event_Dataset
    else
        Reject the article
    end if
end for

Step 4: Extend the *Event_Dataset* by adding more words in the *EventRelatedKeywords* list

Extract keywords and their relevance score from Event_Dataset using **RAKE**
Remove keywords with relevance score less than 20% of the highest score
Remove keywords with more than 2 words
Sort keywords by their relevance score.
Manually annotate top 100 keywords, pick relevant keywords from them and add them in *EventRelatedKeywords* list.

Repeat the whole process Step 1 to 4 until no more relevant keyword found by augmenting *keyword_set* with new keywords every time.

jects.Data returned by the API is in the form of objects, which can be handled as json. While using the API(or tweepy), we may encounter a number of errors, for which twitter returns an error code.Almost all such cases are of rate limiting and unauthorized access.**Rate Limiting:**The calls(functions) provided by the API(or tweepy) are rate limited i.e. in a window of a certain time period, we can only make specified number of calls.**Unauthorized ac-**

Event	Description
Demonetization	demonitis, demonitiz, 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
Goods and Services Tax	GST, Goods and Services Tax, Goods & Services Tax, excise duty
Farmers issues	loan waiver, farmer loan, farmer suicide, pest infestation, Swaminathan Commission, National Commission on Farmer, kisan, monsoon failure, crop failure, fertilizers, Seeds Corporation, farmer, agricultural

Table 3.2: Keywords used to find articles of economic policy issues

cess: Mostly occurs when trying to access tweets of an user who requires you to follow them to check their tweets.Such errors need to be handled through exception handling.

Two types of tweets were collected:

- Tweets made by newspaper community of the dailies under study (called **community tweets**).Given a twitter userid we can download upto about 3000 latest tweets for the user using the **tweepy** function.The downloaded tweets of an user are filtered for a specific event by searching the event-specific keywords in the tweet text and in the hashtags of the tweets.

- Tweets sharing the article urls of policies under study (called **url tweets**). Articles published online by the newspapers are obtained by our Article Extraction Algorithm. These articles are shared on social media (twitter). These URL tweets are needed to be collected for various studies. This is done by scraping the **twitter advanced search** site. The site provides many features to search tweets, one of them being

Event	Description
Aadhaar	Aadhar, Aadhaar, Adhar, Adharcard, Aadharcard, Aadhaarcard, UIDAI, Aadhar Card
Digital India	Digital India,digital india,Digital Swades,digital swades,India Digital,india digital,Digit India,digit india,Digital Desh,digital desh,make 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,Bharat net,digitalised,digitalized
E-Governance	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
Casheless 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 , banking transaction, swiping machine,payment gateway

Table 3.3: Keywords used to find articles of ICTD issues

searching by url shared in the tweet text. To search on the site, we need to formulate a search query. For our purpose the query will be **url**:$< URL >$, where $< URL >$ is the url for which you are searching the tweets for. The scraper used for this is publically available on the internet.

Column	Description
DATE	Date on which the respective questions were asked
SUBJECT	Subject of the Questions asked
QUESTION	Questions asked on the given subject
ApectID1	First Topic(Identified by running LDA on the respective event articles) to which the Questions asked belong
$< event\_name > MassMediaAspect1$	Name of the topic
AspectID2	Second Topic(Identified by running LDA on the respective event articles) to which the Questions asked belong(If Any)
$< event\_name > MassMediaAspect2$	Name of the second topic(If Any)
MINISTRY	ministry of the Lok Sabha who has asked these Questions
MPs	Members of the Parliament who have asked the Questions
ID	Question ID
Q_No	Question Number
Q_Link	Link to the Questions

Table 3.4: Relational database schema: QH data stored

3.2 Data Processing:

3.2.1 Latent Dirichlet Allocation(LDA):

After event specific articles have been collected and stored in MongoDB,we run LDA on the data to extract aspects.We apply preprocessing steps on the articles(removing punctuation, removing stop words, and adding bigrams and trigrams that occur more than 100 times across the corpus). Then each document is converted to Bag of Words form, in the form of (word-id, word count in document).Then the LDA models are generated, using varying values for number of aspects.We check from 5 aspects to 34 aspects.The coherence values are calculated for each, and through pyLDAvis tool we visualize different LDA models for different number of aspects.The one model with the best

separation of aspects and high coherence value is selected-this involves visually checking the pyLDAvis plots, and looking at the top words and random documents from each aspect to see if the generated aspects actually correspond to separate coherent topics. Finally,we map the documents to different aspects.A document is mapped to an aspect if the probability of the document belonging to that aspect is greater than or equal to 0.3.We also write out a file with 50 words per aspect with their probability values and top 20 documents for each aspect.

After this mapping is done,we manually name all the aspects and we also merge all the common aspects.We extracted 19 aspects for Aadhaar,15 aspects for Demonetisation,11 aspects for GST and 14 aspects for Farmer's Protest. We then randomly select some articles of a particular event and manually annotate those articles to different aspects to check the accuracy of our LDA methodology.The accuracies were 96% for Aadhaar,85% for Demonetisation,70% for GST and 76% for Farmer's Protest.

3.2.2 Sentiment Extraction:

We use SentiStrength [45] and Vader [44] tool for extracting sentiments of articles and tweets respectively. The SentiStrength tool gives us two values:TPOS and TNEG.The TPOS values range from 1(least positive) to 5(most positive).The TNEG values range from -1(least negative) to -5(most negative).We compute the compound sentiment of an article as the summation of TPOS and TNEG.We compute the degree of polarity(degpol) of an article using the following formula where 'a' refers to an article:

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

The Vader toot gives us four values.It gives us a pos(positive component of the tweet),neg(negative component of the tweet),neu(neutral component of the tweet) and compound(overall sentiment of the tweet).The polarity value of a tweet is computed by the same formula which is mentioned for an article.

3.2.3 Entity Extraction and Entity Resolution:

We use *Open Calais* tool to extract entities out of the article texts and tweets.We extract entities mainly of type Person,Company,Organisation,City etc.We also store the relevance scores of each entity extracted. Before doing any entity based coverage bias or sentiment bias analysis,we resolve the entities extracted against each other using their context information.For example,if in some article we find **Narendra Modi** and in another article we find **Prime Minister**,resolving these two entities means identifying them as the same person,merging them and saving **Prime Minister** as the context information for **Narendra Modi**.

3.3 Database and Our Analysis Pipeline:

3.3.1 Creating the master tables:

We have our media database(postgres database) in our server.We have the following master tables:

- event_info
- aspect_info
- lda_info
- art_info_final
- full_tweets_dataset
- direct_tweets_dataset
- community_tweets_dataset

event_info:

This relation stores all the information about a particular policy we are trying to analyse.The schema of the relation is shown in table 3.5:

Column	Description
event_id	The id of a policy
event_name	The name of the policy
keywords	Keyword set used to extract articles of the policy
description	Description of the policy

Table 3.5: Relational database schema: Event Info stored in database

Column	Description
event_id	The id of a policy
aspect_id	The id of an aspect for a policy
aspect_name	The name of the aspect
description	Description of the aspect

Table 3.6: Relational database schema: Aspect Info stored in database

Column	Description
event_id	The id of a policy
asp_id	The id of an aspect for a policy
keywords	Top keywords for every aspect
prob	Probability values of those keywords

Table 3.7: Relational database schema: LDA Info stored in database

aspect_info:

This relation stores all the information about the aspects extracted for a particular policy we are trying to analyse.The schema of the relation is shown in table 3.6:

lda_info:

This relation stores all the information about the top keywords and their probabilities for different aspects identified for a particular policy through LDA.The schema of the relation is shown in table 3.7:

Column	Description
event_id	The id of an event
article_id	Id of an article taken from our MongoDB
asp_id	Array contain the aspect ids to which an article belongs
category	Category of the article
url	Url of the article
word_cnt	Number of words in the article
newssource	Newspaper which has written the article
senti	Overall sentistrength of the article
senti_pos	Positive sentistrength of the article
senti_neg	Negative sentistrength of the article
vader_comp	Overall Vader sentiment of the article
vader_pos	Positive Vader sentiment of the article
vader_neg	Negative Vader sentiment of the article
date	Publishing date of the article

Table 3.8: Relational database schema: Article Info stored in database

art_info_final:

Once the articles for an event are obtained and stored in our MongoDB collection and LDA,sentiment extraction,entity extraction and resolution have been completed,we push in all the article data into **art_info_final** relation in our database.The schema of this relation is shown in table 3.8:

full_tweets_dataset:

We store the information of all the tweets that have been collected for a policy into this relation.The schema of this relation is shown in table 3.9:

direct_tweets_dataset:

We store the information of all the direct URL tweets that have been collected for a policy into this relation.These are the tweets which have been obtained by hitting the article urls.The schema of this relation is shown in table 3.10:

Column	Description
event_id	The id of an event
tweet_id	Id of the tweet
compound_sent	Overall sentiment of the tweet
positive_sent	Positive sentiment of the tweet
negative_sent	Negative sentiment of the tweet
text	Text of the tweet
hashtags	Hashtags in the tweet
retweet_count	Number of retweets of the tweet
user_id	Id of the user who made the tweet
newspaper	Newspaper list the user follows
tweet_data_time	Date and time of the tweet

Table 3.9: Relational database schema: General Tweets Info stored in database

Column	Description
event_id	The id of an event
tweet_id	Id of the tweet
compound_sent	Overall sentiment of the tweet
positive_sent	Positive sentiment of the tweet
negative_sent	Negative sentiment of the tweet
url	URL mentioned in the tweet
text	Text of the tweet
hashtags	Hashtags in the tweet
retweet_count	Number of retweets of the tweet
user_id	Id of the user who made the tweet
newspaper	Newspaper list the user follows
tweet_data_time	Date and time of the tweet

Table 3.10: Relational database schema: Direct URL Tweets Info stored in database

community_tweets_dataset:

We store the information of all the community(follower) tweets that have been collected for a policy into this relation.These are the tweets which have been made by the followers of different newspapers.The schema of this relation is shown in table 3.11:

Similar **master tables** for mass media are also maintained for **Hindi arti-**

Column	Description
event_id	The id of an event
tweet_id	Id of the tweet
compound_sent	Overall sentiment of the tweet
positive_sent	Positive sentiment of the tweet
negative_sent	Negative sentiment of the tweet
url	URL mentioned in the tweet

Table 3.11: Relational database schema: Follower Tweets Info stored in database

Column	Description
event_id	The id of an event
aspect_id	Aspect Id
newspaper	Name of the news source
coverage	Relative coverage given to a particular aspect
agg_sent	Aggregate Sentiment expressed in the articles for the aspect by the newspaper
polarity	Degree of polarity

Table 3.12: Relational database schema: Result of newspaper wise coverage and sentiment bias

cles.

3.3.2 Analysis Pipeline and the Resulting Tables:

Once we have all our master tables populated,we perform coverage bias, sentiment bias and various other analysis.We perform the following analysis one by one:

- **Newspaper wise coverage and sentiment bias analysis in mass media**: We try to find out how much coverage different news sources are giving to different aspects of a policy and how are they covering those aspects(sentiment bias).We store the result as shown in table 3.12:
- **Coverage Bias of different aspects for a policy across newspapers:** We try to see how the aspects are covered across newspapers for

Column	Description
event_id	The id of an event
aspect_id	Aspect Id
massmedia_coverage	Relative coverage given to a particular aspect in mass media
socialmedia_direct_coverage	Relative coverage given to a particular aspect in the direct URL tweets
socialmedia_fol_coverage	Relative coverage given to a particular aspect in the follower tweets
qh_coverage	Relative coverage given to a particular aspect in the Question Hour data

Table 3.13: Relational database schema: Coverage Bias across different media over all newspapers

Column	Description
event_id	The id of an event
newspaper	News source
symmetric_kl_div	How much the news source diverges
mean_entropy	Mean Entropy Values

Table 3.14: Relational database schema: How much a newspaper diverges from mean aspect coverage

a particular policy and compare the results across mass media,social media and QH data.We store the result as shown in table 3.13:

- **KL Divergence:** We try to find out how different news sources diverges from mean aspect coverage for a particular policy.We store the results as shown in table 3.14:

- **Coverage bias of different newspapers towards different constituencies:** We try to find out how much coverage different news sources give to different constituencies like poor,middle-class,corporate, informal sector and government.We store the results as shown in table 3.15:

- **Comparison of Coverage given to different constituencies across different media:** We try to compare the coverage bias towards constituencies across mass media,social media and QH data.We store re-

Column	Description
event_id	The id of an event
newspaper	Name of the news source
category	Name of the constituency
coverage	Relative coverage given to a particular constituency

Table 3.15: Relational database schema: Coverage bias of different newspapers towards different constituencies

Column	Description
event_id	The id of an event
category	Name of the constituency
massmedia_coverage	Relative coverage given to a particular constituency in mass media
socialmedia_fol_coverage	Relative coverage given to a particular constituency in the follower tweets
qh_coverage	Relative coverage given to a particular constituency in the Question Hour data

Table 3.16: Relational database schema: Comparison of Coverage given to different constituencies across different media

Column	Description
event_id	The id of an event
aspect_id	Aspect Id
aspect_cov	Relative coverage given to a particular aspect across newspapers
aspect_aggr_sentiment	Total sentiment given to an aspect across newspapers
polarity	Degree of polarity

Table 3.17: Relational database schema: Aspectwise Coverage and Sentiment Bias of the opinion articles for a particular policy

sults as shown in table 3.16:

- **Aspectwise Coverage and Sentiment Bias of the opinion articles for a particular policy:** We try to find out how much coverage is given to different aspects for the opinion articles across newspapers.Results are stored as shown in table 3.17:

Column	Description
event_id	The id of an event
aspect_id	Aspect Id
newspaper	Name of the news source
coverage	Relative coverage given to a particular aspect
agg_sent	Aggregate Sentiment expressed in the articles for the aspect by the newspaper
polarity	Degree of polarity

Table 3.18: Relational database schema: Newspaper wise coverage and sentiment bias for opinion articles

Column	Description
event_id	The id of an event
author_name	Name of the author
url_list	The list of urls of the opinion articles written by the author
word_cnt_list	The list of word counts of the opinion articles written by the author
newspaper_list	The list of news sources of the opinion articles written by the author
agg_sent_list	The list of overall sentiments of the opinion articles written by the author
pos_sent_list	The list of positive sentiments of the opinion articles written by the author
neg_sent_list	The list of negative sentiments of the opinion articles written by the author

Table 3.19: Relational database schema: Relation storing the meta info of the authors who have written the opinion articles

- **Newspaper wise coverage and sentiment bias for opinion articles:** We try to find out how much coverage different media houses give to different aspects and how they cover them for the opinion articles.We store the results as shown in table 3.18:

- **Relation storing the meta info of the authors who have written the opinion articles:** We try to compute and store the meta info about all the authors who have written the opinion articles.Results are stored as shown in table 3.19:

Column	Description
event_id	The id of an event
aspect_name	Name of the author
count_opeds_written	Number of opinion articles written by the author
agg_sent	Aggregate Sentiment expressed in the opinion articles written by the author
degpol	Degree of polarity

Table 3.20: Relational database schema: Sentiment Bias for the authors who have most frequently written the opinion articles for a policy

Column	Description
event_id	The id of an event
event_name	The name of an event
power_elite_or_not_1_or_0	Status of the whether the entity is a power elite or not
role	Role of the entity
total_about	Total number of statements made about the entity
total_by	Total number of statements made by the entity
category	category of the mass media articles we have analysed

Table 3.21: Relational database schema: Coverage bias for the entities covered in mass media

- **Sentiment Bias for the authors who have most frequently written the opinion articles for a policy:** We try to find out the sentiment bias of the authors who have written most number of opinion articles for a policy.We store the results as shown in table 3.20:
- **Coverage bias for the entities covered in mass media:** We try to find out how much coverage is given to different entities like politicians, academicians,judiciary in mass media.We store the results as shown in table 3.21:
- **How have the entities spoken in mass media:** We try to find out how have the entities spoken in mass media.We analyse the by

Column	Description
event_id	The id of an event
event_name	The name of an event
aggr_sent	Aggregate sentiment expressed by the entity
tot_pos	Aggregate positive sentiment expressed by the entity
tot_neg	Aggregate negative sentiment expressed by the entity
polarity	Degree of polarity
category	category of the mass media articles we have analysed

Table 3.22: Relational database schema: How have the entities spoken in mass media

Column	Description
event_id	The id of an event
event_group	The name of the entity group
total_about	Total number of statements made about the entity
about_percentage	Percentage of statements made about the entity
total_by	Total number of statements made by the entity
by_percentage	Percentage of statements made by the entity
category	category of the mass media articles we have analysed

Table 3.23: Relational database schema: Coverage bias for the entity groups in mass media

statements made by the entities.We store the results as shown in table 3.22:

- **Coverage bias for the entity groups in mass media:** We try to find out how much coverage is given to different entity like bjp,inc, corporate in mass media.We store the results as shown in table 3.23:

All these analysis use the **master tables** from our database as their inputs.All the results are also stored in our database in the format as described above.This is the entire Analysis Pipeline that we are having for our Media

Analysis.Some of the results are also uploaded on our Media Website,the structure of which is explained in the next section.

3.4 Media Website:

We are working towards a vision to build a data driven economy monitoring engine that will bring together economic time-series, longitudinal satellite data, and media reports, to improve the awareness of citizens on how economic policies get shaped, their effects, and give an overall state of the socio-economic lives of people as a result of policies and actions implemented in the Indian democratic framework. Our aim is to build a first version of GEM (Giant Economy Monitor) in India.

How economies work is poorly understood by most citizens how corporate and political interlocks shape industrial policies or resource allocation, how political alignment of elected representatives affects the growth of their constituencies, how local and national inequality is changing over time, the impact of government welfare schemes, level of policing of trade malpractices such as commodity hoarding or price manipulation, how media ownership influences its bias, whether social media helps level mass media bias or amplifies echo chambers, are aspects that if understood by citizens can help strengthen the democratic frameworks to make the world a better place for everybody. Representing the functioning of the economy in simple terms in hard data collection from many sources, data mining to detect relationships based on state of the art in economic/financial/political models, and making it accessible to people in contextual ways is what we aim to achieve.

Different teams working in different domains(media bias analysis,analysing corporate and political interlocks,satellite data analysis) upload their findings and results onto our media website under different tabs.We list out the different tabs we have on the Home Page of our website:

- **Corporate Government interlocks:**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 social network that becomes a power structure of highly influential entities. This power structure often leads to a bidirectional flow of favors between the corporations and political entities, and it can influence policy formulation or manifest itself in crony capitalism (or corruption). Such negative manifestations of corporate-government interlocks have been evident in India (for example, in 2G scam, where an interconnected group of politicians and corporate business-persons were accused of cronyism).

Under this tab,the results of the analysis done on corporate-government interlocks are shown.

- **Political Economy:**As defined by 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. Thus, in simple terms, it studies the interaction between corporates (actors involved in production and trade), government, and bureaucrats (responsible for law, custom, and governance). Among the influential actors who are somehow involved in the policy making process, or are involved in the policy discourse, we include actors like politicians, business-persons, bureaucrats, social activists, academics, policy experts, and judiciary members. We call these influential actors elites, and study what they speak about and how they speak about, with respect to a policy.

  We have done analysis on four related events like digital india, e-governance,cashless economy and aadhaar.On clicking a particular event, we can see which entities or entity groups were more vocal in mass media (entity coverage) in terms of the statements they made about a policy, and the aggregate sentiment slant of their statements about a policy. This analysis is done across all of the seven newspapers considered in the media space.

- **Media Bias:**Alignment of newspapers towards the five constituencies of poor, middle class, informal sector and small trades, corporate, and government. Using the plot on the right, we try to answer the research

question: "Do news-sources have a bias towards or against constituencies like the poor, middle class, corporate, informal sector, and government?". Through this research question, we try to analyse the effect of framing in mass media. Framing refers to the modes of presentation that media houses to present information in a way that aligns with the readers underlying schemas 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 analyse this effect by automatically extracting aspects (topics of discussion) from the news articles, and manually linking them with the five constituencies of poor, middle class, informal sector and small trades, corporate, and government. For each policy, we map each aspect to the constituencies to which it is relevant, and also see the alignment of the aspect towards the constituency in terms of the sentiment of its constituent articles.

- **Socio-Economic development:**Models for socio-economic development are useful for planners to build appropriate policies. Such models are ideally constructed based on empirical data, and the concerned team has taken up the problem of working towards a district development model for India by using two waves of census data. India has almost six hundred districts, diversely spread apart in terms of both social and economic development, and hence presents a unique natural experiment to understand how social and economic factors interplay with one another. They present some interesting observations that they are able to make from longitudinal analysis of census data from the years 2001 and 2011, and raise some questions calling for the need for ethnographic and other surveys to be able to understand the underlying mechanisms that would lead to the observed patterns.Some of these mechanisms can be captured using data from participatory media networks, and propose as part of future work to use this bottom-up data together with top-down data from censuses and other sources, to build a comprehensive district development model.

Chapter 4

Media Bias Analysis on four economic policies of India

4.1 Data Collection:

We cover four main events in this analysis.We cover **Aadhaar, Demonetisation, GST and Farmer's Protest**.Articles and Tweets for these events are collected using the algorithm mentioned in **Chapter 3**.We collected **12809** articles for **Aadhaar**,**17849** articles for **Demonetisation**,**15756** articles for **GST** and **13840** articles for **Farmer's Protest**.In order to do a newspaper based analysis for the tweets,we took the tweets which were made by the followers of different newspapers containing the article urls quoted in the tweet text.We collected **34521** tweets for **Aadhaar**,**59489** tweets for **Demonetisation**,**38073** tweets for **GST** and **22820** tweets for **Farmer's Protest**. After data collection,aspects are extracted using LDA(as explained in Chapter 3).Using Vader and SentiStrength tool,sentiments for articles and tweets are computed and entities are extracted using open calais tool.

4.2 Methodology:

Methodology used like articles and tweets extraction,LDA algorithm and sentiment and polarity computations,all have been explained in details in the **Data Collection** and **Data Processing** sections in Chapter 3.

4.3 Results:

In this section, we answer the first two research questions mentioned in the Introduction, and present the relevant results.

4.3.1 (RQ1:)Are news-sources biased on the amount of coverage they give to different aspects about the policy issues?

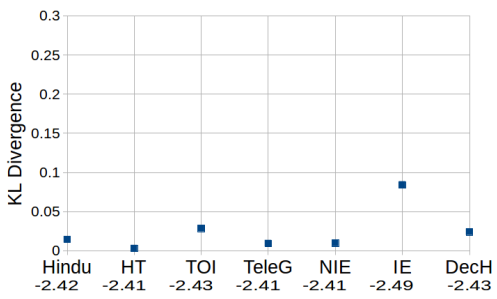
In this research question,we compare the coverage different news sources give to different aspects about these policies.We define the relative aspect coverage for different news-source as follows:-

$$relative\_aspect\_coverage = \frac{count(words,aspect)}{\sum_{aspect}(words,aspect)}$$

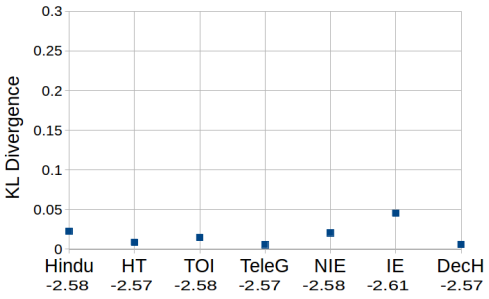
where count(words,aspect) is the total number of words of all the articles belonging to the aspect published by the news-source.To see how much coverage a particular news-source give to different aspects,we first compute the mean aspect coverage by simple averaging the relative coverage given to an aspect by all the news-sources.We then see how much a particular news-source diverge from the mean aspect coverage by plotting the KL divergence of the two distributions (relative aspect coverage distribution for a news source, and the global relative aspect coverage distribution across news-sources) for each news-source as shown in Figure 4.1.

For all the events we have covered,we see that Indian Express(IE) which is believed to be a pro-INC newspaper has the maximum deviation from the mean aspect coverage compared to other news sources.This means that it has covered aspects differently compared to other news sources.It has given coverage to such aspects which have not much been covered by other news sources.Moreover,we see that apart from IE,all the other news sources are having similar deviations(divergence) from mean aspect coverage for *Aadhaar,Demonetisation and GST*,which means all the news sources have given similar coverage to different aspects for these policies.We see Indian Express to cover more political aspects and issues like they give a huge coverage to *Parliamentary debates on Aadhaar* in Aadhaar policy event,aspects like *Appreciation by PM for supporting Demonetisation and Court cases related to Aadhaar(right to privacy and social welfare schemes)* for demonetisation.In *Farmer's Protest*,we see that all the news sources have significant variations in their divergence from mean aspect coverage.In *Farmer's Protest*,we see Dec-

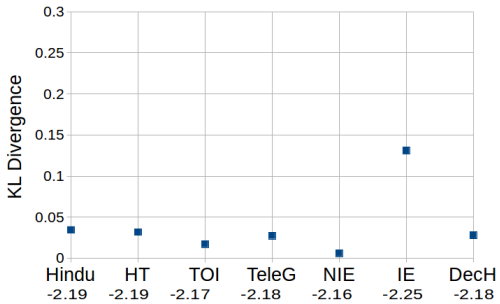
Figure 4.1 [RQ1] KL divergence across aspects between distributions of relative coverage and mean relative coverage (across news-sources) for the four events. Higher the deviation for a particular news source, more different is its coverage of aspects from the mean behavior across news-sources. Values on X- axis represent $\sum p * \log m$, where p represents the probability distribution of aspects for a news-source, and m represents the mean distribution of aspects across news sources.



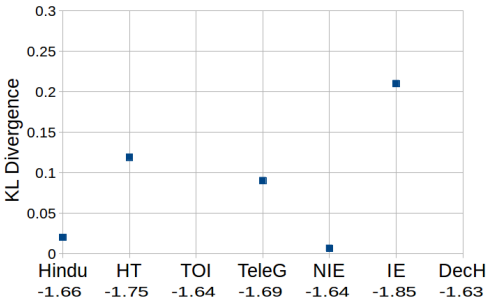
(a) Demonetisation



(b) Aadhaar



(c) GST



(d) Farmers' Protest

can Herald and Hindustan Times covering more political aspects like *Protests by farmers and Crimes and Suicide in farmer community.*On the other hand we see Hindu and Indian Express giving the most coverage to implementation oriented aspects like *Loan waiver implementation by State Governments.* We also show the relative coverage given by different news sources to different aspects in mass media about Aadhaar in section A.4,about Demonetisation in section A.5,about GST in section A.6 and about Farmer's Protest in section A.7.

4.3.2 (RQ2:)Are some news-sources more closely aligned with their readers (on social media) than others?

In this research question,we see whether the news sources are aligned with their readers in social media(twitter).Often the readers of different newspapers tweet the url of the article with some short comment.We consider the user tweeting an article url to be the reader of the corresponding newspaper.We consider all the retweets as well.We analyse how much the mean sentiment of all the tweets about the articles of a news source align with the sentiment of its news articles.Let us define two terms:-

- $S(a, n)$: This refers to the sentiment of an article 'a' for news source 'n'
- $S(t, a, n)$: This refers to the mean sentiment of all the tweets containing the url of article 'a'.

We the compute two vectors:-

- **V1**: This contains $S(a, n)$ values for all the articles of news source 'n'.
- **V2**: This contains $S(t, a, n)$ values for all the articles of news source 'n'.

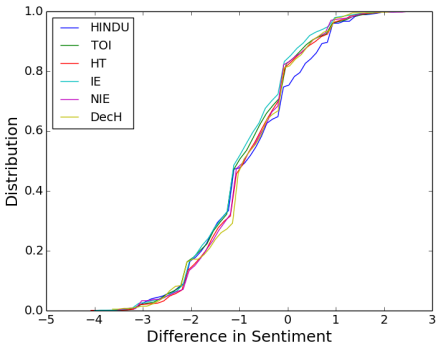
We then do paired T-tests between V1 and V2 for every news source.Results are shown in table 4.1.We see that the p-values for all the news sources are very low which indicates strong alignment between the news source and their

News Source	Demonetisation		Aadhaar		GST		Farmers' Protest	
	T-statistic	P-value	T-statistic	P-value	T-Statistic	P-value	T-statistic	P-value
Hindu	-12.659	1.139E-31	-4.871	1.358E-06	-12.312	1.527E-30	-20.585	1.407E-74
HT	-23.924	7.384E-104	-10.744	2.904E-23	-17.081	9.752E-55	-28.706	5.471E-116
IE	-37.074	1.36E-232	-12.317	1.101E-31	-17.801	3.407E-61	-34.408	8,273E-172
NIE	-20.384	7.234E-76	-7.393	1.521E-12	-10.469	2.382E-23	13.574	1.82E-28
TOI	-37.824	3.34E-253	-21.411	1.463E-93	-27.787	2.343E-149	-42.884	1.204E-280
DecH	-13.92	2.736E-36	-8.155	3.363E-15	-10.341	6.688E-23	-16.317	3.35E-41

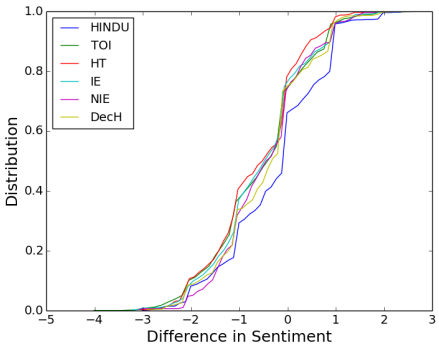
Table 4.1: [RQ2] Paired T-test on (article sentiment - Mean tweets sentiment). $p < 0.001$ shows both distributions are similar at 99.9% significance.We could not collect any tweets containing article urls of telegraph.

readers.We plot the (V1-V2) distribution in a CDF plot for every news source as shown in figure 4.2.We also see a steep rise at 0.0 value which strongly validates our argument of strong alignment.We see that Times Of India(TOI) has the maximum alignment with its readers(audience) followed by Indian Express(IE) and Hindustan Times(HT).This shows that the audience are ideologically aligned with their news source and are different from each other in their beliefs.We could not collect any tweets containing article urls of telegraph.

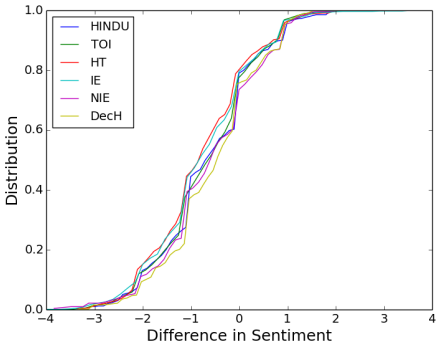
Figure 4.2 [RQ2] CDF plot of (Article Sentiment - Mean Tweets Sentiment) i.e. $V1 - V2$ for four policy events.



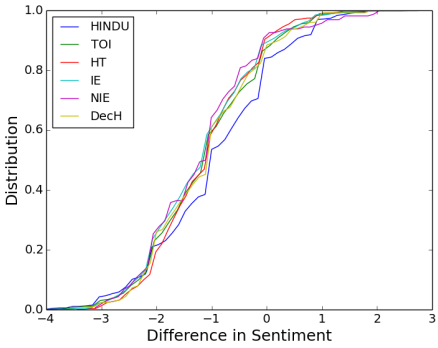
(a) Demonetisation



(b) Aadhar



(c) GST



(d) Farmers' Protest

Chapter 5

Analysis of Parliamentary Question Hour Data, Mass Media Data, and Social Media Data to Study the Discourse on Policies

5.1 Methodology:

5.1.1 Collection of articles,tweets and QH data:

Collection of articles from a set of selected keywords and collection of tweets from the follower handles of different media houses have been explained in details in Chapter 3.LDA was applied on the articles that we collected and aspects were extracted.After that every article was mapped to at least one aspect.Details of the algorithm and the number of aspects that we extracted per event can be found in Chapter 3.Since we needed to do an aspect wise analysis on the follower tweets,we considered only those tweets which had the article urls quoted.We collected **12809** articles for **Aadhaar**,**17849** articles for **Demonetisation**,**15756** articles for **GST** and **13840** articles for **Farmer's Protest**.We collected **34521** tweets for **Aadhaar**,**59489** tweets for **Demonetisation**,**38073** tweets for **GST** and **22820** tweets for **Farmer's Protest**.

We collect the question hour (QH) data from the website for Lok Sabha, and study it at two levels: (a) by studying the questions relevant to each policy event separately, and (b) by studying the entire QH dataset that includes questions asked during the 15th (2009-1014) and 16th (2014-till date) Lok Sabha terms (not restricted to the four policies considered). The Lok Sabha" is the term used to refer to the Lower House of the Indian parliament.QH data specific to an event are collected using the same set of keywords which

were used to extract the articles.

5.1.2 Aspect Identification using LDA:

The entire algorithm of LDA and how the aspects were extracted from articles have been explained in Chapter 3.Since the tweets in social media are very short,so LDA was not applied on the tweets.Therefore,for doing an aspect wise analysis on the tweets,we consider only the tweets where an article url was quoted.Next, 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.Therefore,we are able to compare mass media,social media and QH data on same aspects.

5.1.3 Mapping Aspects to Constituencies:

In one of our research questions,we try to analyse the coverage given to different standard constituencies to study the framing in mass media.We identify five constituencies:

- **Poor:**This is the constituency who typically benefit from the wealth distribution strategies.
- **Middle-Class:**This is the constituency having a disposable income and who benefit from tax breaks and lower market prices.
- **Corporate:**This includes big corporate organisation dealing in economic growth,free market policies.
- **Informal Sector:**This includes small enterprises aided by slow formalisation of industries and trade.
- **Government:**This includes pro/anti view points towards a state.

We then map each of the aspects identified for each policy to different constituencies with **+1/-1/0** values. A +1 value for a constituency means that the aspect is pro-constituency and a -1 value means that the aspect

is anti-constituency.A 0 value means that the aspect is not related to the constituency.For example,in Demonetisation we have an aspect named **Long Queues at ATM/Banks and Cash Crunch**.Now we found most of the articles for this aspect to be negative,i.e pro-middle class since it is talking about the problems of common people.Same aspect was mapped as anti-government because the articles were condemning the policy. The details of aspect to constituency mappings for different policies can be found in Appendix under section A.3.

5.2 Results:

5.2.1 (RQ3:)Which aspects and constituencies about the policies do the mass media,social media, and QH data cover?

In this research question,we try to find out which aspects are covered across newspapers for mass media,social media and QH data.To compute the relative coverage for aspects in mass media,we use the following formula:

$$relative\_aspect\_coverage\_massmedia = \frac{count(words,aspect)}{\sum_{aspect}(words,aspect)}$$

where count(words,aspect) is the total number of words of all the articles belonging to that aspect across all the news sources.

To compute the coverage of tweets,we use the following formula:

$$relative\_aspect\_coverage\_socialmedia = \frac{count(tweets,aspect)}{\sum_{aspect}(tweets,aspect)}$$

where count(tweets,aspect) is the total number of tweets belonging to that aspect.

To compute the coverage of QH data,we use the following formula:

$$relative\_aspect\_coverage\_QH = \frac{count(questions,aspect)}{\sum_{aspect}(questions,aspect)}$$

where count(questions,aspect) is the total number of questions mapped to that aspect.

Next we present our analysis to see which aspects get what coverage in mass media,social media(follower tweets) and Question Hour data.We present our analysis on an event basis.Our main aim to analyse aspect coverage in social media is to see whether there exists any ideological differences between the news sources and their audience.We find out that the aspect coverage trend in social media is similar to that of mass media for all the policies.So,we mainly compare the differences between mass media and Question Hour data.

Aadhaar: This issue has been criticized owing to lack of security and privacy in citizens data collection and storage mechanisms, and also because of an allegedly faulty implementation of the platform or use of the platform by different agencies.In mass media,we find that more coverage is given to legal issues like **court cases related to aadhaar** and some practical issues like **problems with the enrollment centers**.The QH data also talks more about similar issues.

The highest covered aspects in mass media talks about problems with enrollment centers and court cases related to aadhaar which primarily deals with data security,right to privacy and linkage of aadhaar to welfare schemes.This aspect of court cases related to aadhaar mainly discusses the court cases on whether the right to privacy is a fundamental right,and the issues on leakage of personal data of citizens due to lack of planning and professionalism.This aspect mainly focuses on the issues of middle class and there has been mentions in the articles about middle class being the most affected constituency.The aspect which talks about the technical problems related to the implementation of UIDAI scheme and problems with enrollment centers has been covered the most.This too primarily refers to middle class as it talks about the problems faced by common people during aadhaar registration or long queues at enrollment centers,etc..When there is some criticism or talks about any policy,the government constituency also gets covered a lot.So we find out that **middle class and government** are the two most covered constituencies.

For QH data, **Aadhar Enrolment centers and UIDAI operations** 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. While this aspect also discusses issues like setting up of Aadhaar enrollment centers in metro cities,the questions are more related to middle class of the society than poor.The second highest covered aspect in QH data **Problems with Aadhaar card cancellations and enrollment centres** discusses the steps taken by the government against the loopholes in the policy implementation. The issues discussed include cancellation of thousands of Aadhaar cards, the reasons behind cancellation, insufficient number of enrollment centers set up,etc.This aspect again does not explicitly include questions on issues of the poor, and is more oriented towards the reprimands issued to the government owing to faults in implementation of the policy.So this aspect mainly targets the government constituency.So we see that both in mass media and QH data,more importance has been given to the middle class section compared to poor.

Demonetisation: The move was widely criticized owing to multiple problems caused to common people due to sudden depletion of liquidity, irregularities in norms of exchanging old currency notes, cash exhaustion in ATMs, etc. The suddenness of the policy move also led to suffering on part of the farmers and daily wage earners due to shortage of cash among people, and understaffed banks unable to dispense cash.
In mass media we find out that **Appreciation by PM for supporting Demonetisation** is the highest covered aspect.This aspect mostly talks about the prime minister's statements thanking the common people for supporting the policy move to fight black money and corruption.In this aspect,the articles mainly talked about the expression of gratitude,the prime minister showed towards the citizens for supporting this policy,encouraging them,thanking them for bearing the troubles faced post demonetisation and explaining them the benefits it will bring in the long run.The second most covered aspect is **Opposition unites against government on Demonetisation**.This aspect mainly talked about the political blame games around politicians,the comments from the opposition protesting demonetisation and statements from the prime minister calling the people from opposition as

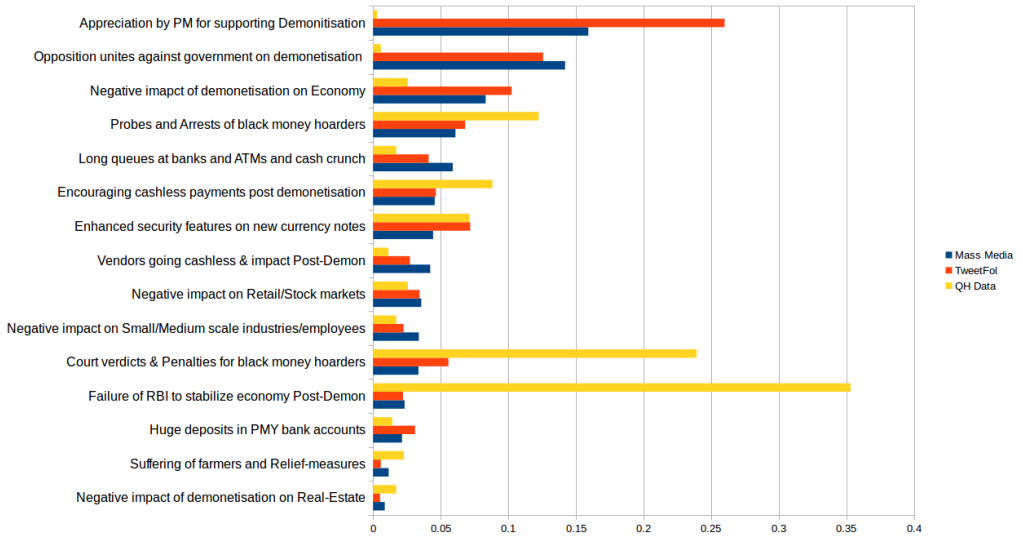
anti-poor and anti-middle class.
In QH data,we find that the most covered aspects are quite different from the mass media.The highest covered aspect in QH data is **Failure of RBI to stabilize economy post demonetisation** which included questions on issues like the menace of fake currency that circulated immediately after the move and the issues related to currency management.The questions that were asked were mainly asked to the top ministries and The Reserve Bank Of India(RBI) about the steps taken to restore normalcy,amount of fake currency confiscated,etc.The second most covered aspect was **Court verdicts and penalties for black money hoarders** which mainly covered questions like the steps taken against the black money hoarders,their arrests,coordination with Switzerland to detect the graft money deposited in the Swiss Banks,etc. So,in the aspects mostly covered in mass media,we saw that political issues got covered the most.So the articles written were all around the issues of poor and middle class,whereas in the QH data we saw that most questions were about the feedback to the policy,questions on improving the situation post the policy move,etc.So government was the main talking point in the questions asked and nothing much was asked related to poor and middle class.

GST: Since the implementation of this policy,there have been intense debates on its complexity and problems in implementation which have impacted the overall growth of the economy.Both in mass media and QH data,we see that aspects related to implementation of the policy,issues about small traders have been covered the most and not much coverage has been given to problems faced by poor people.
In mass media,the highest covered aspect is **Discussion of GST Bill in the Parliament**.This aspect covered all the political debates and discussions held in the parliament before and after passing the GST Bill since it was a matter of concern for both central and state governments.Next,we see that **Troubles faced by small traders to meet the GST registration deadline** is the second highest covered aspect.This aspect directly discussed about the troubles faced by small traders and businesses across the country in registering for GST because of impending registration deadlines and fault in the GST registration software.The issues faced by the con-

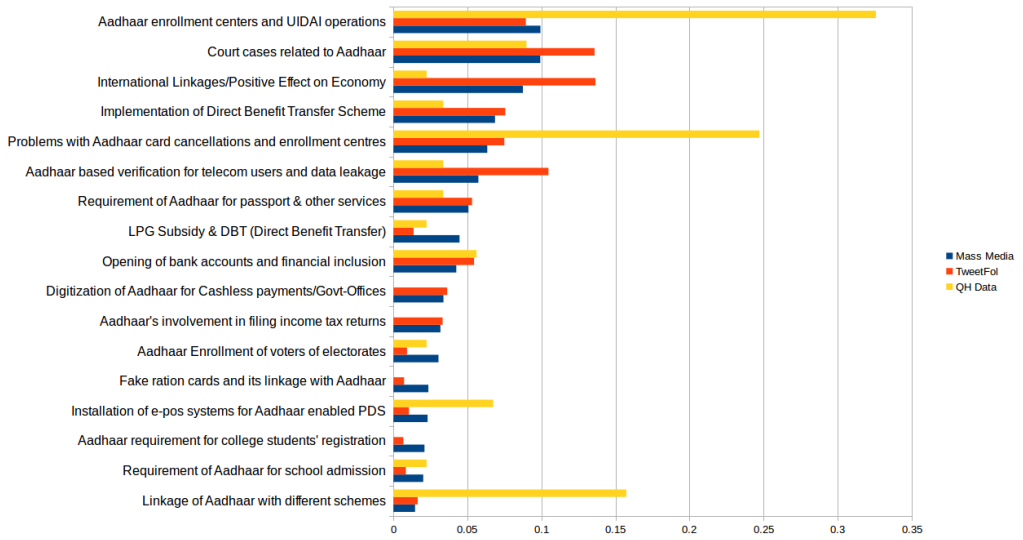
sumers were not taken up and discussed at all. For QH data,the highest covered aspect is same as that of mass media,but apart from that the aspect coverage is very different.The second most covered aspect in QH data is **Revolts/Confusions with GST Slabs**.This includes questions about applicability of GST to various sectors,as the parliament asked most questions related to small traders.Most questions relate to different rates of GST applied to different sectors of trade and none of the questions were about the poor and middle class.In the other highly covered aspects,questions were mostly about the implementation process of GST and the loses incurred by different states because of GST.
So mostly in mass media,issues about small traders and small businesspersons were considered who belong to the middle class constituency.Apart from that in QH data,we see that lots of questions were hurled at the government regarding clearing the confusions about the implementation of the policy and were criticized even regarding its imperfect implementation.

Farmer's Protest: We find in both mass media and QH data,aspects covering loan waiver to farmers were covered the most.The mass media mainly focused on the reason behind the loans taken by farmers and what changes need to be incorporated in the agricultural sector,whereas the parliament mainly asked questions about how to implement loan waiver.
In mass media,the highest covered aspect is **Disbursement of loans and subsidies to farmers by banks**.This aspect included discussions about loans provided to farmers by banks under different government schemes.The second most covered aspect is **Irrigation concerns and water pollution affecting farming** which included discussions about the problems faced by farmers in irrigation, due to water pollution through industrial effluents.
in QH data,the top two covered aspects are **Loan waiver implementation by state governments** and **Disbursement of loans and subsidies to farmers by banks**.The questions asked were mainly technical,mainly around how the loan waiver is to be implemented,whereas in mass media the articles mainly focused on the reasons for these loans taken and changes to be brought to agricultural sector. The plots regarding relative aspect coverage across different media can be seen in Figure 5.1.The plots regarding constituency coverage can be seen in Figure 5.2.

Figure 5.1 [RQ3] Relative aspect coverage of each policy by mass media, social media, and QH data.

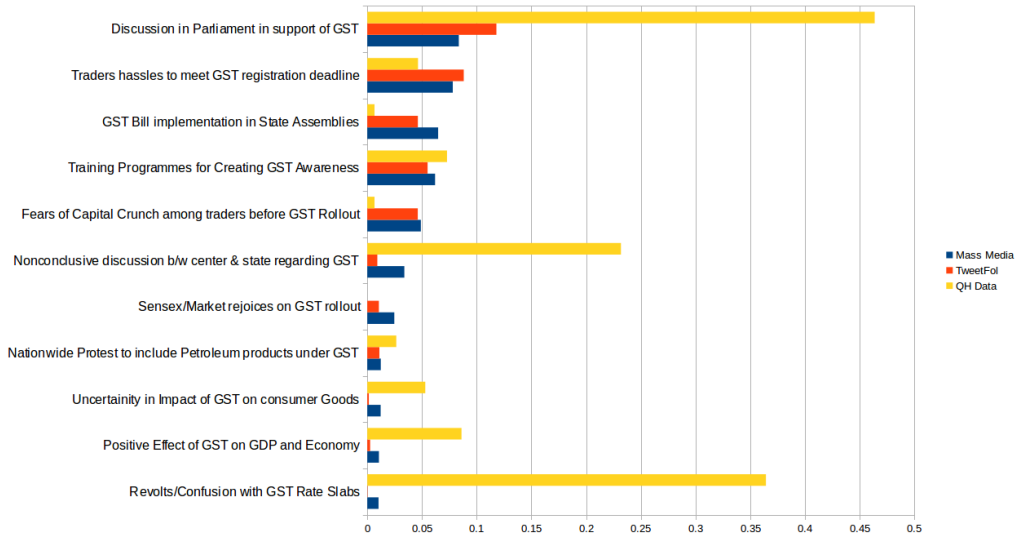


(a) Demonetisation

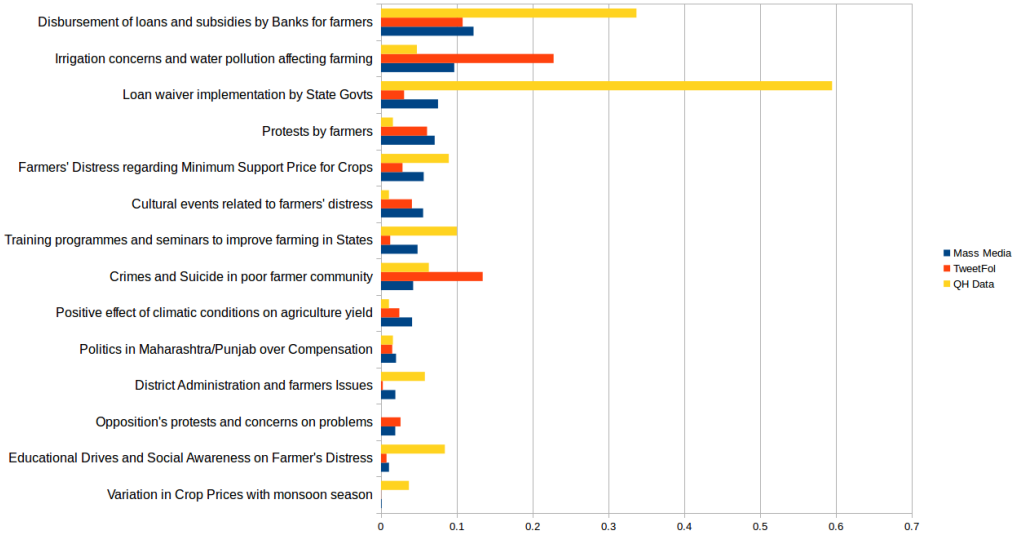


(b) Aadhaar

Figure 5.1 [RQ3] Relative aspect coverage of each policy by mass media, social media, and QH data.

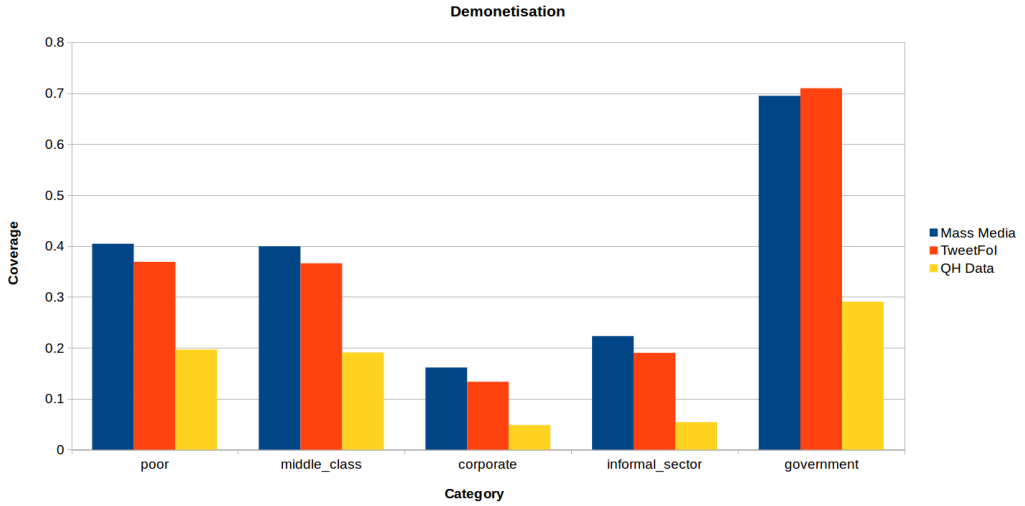


(c) GST

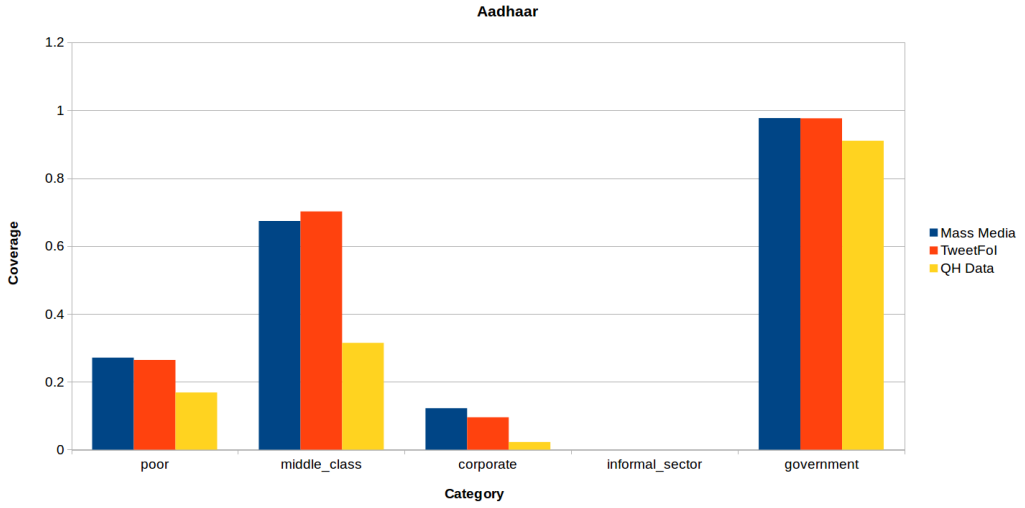


(d) Farmers' Protest

Figure 5.2 [RQ3] Relative constituency coverage of each policy by mass media, social media, and QH data.

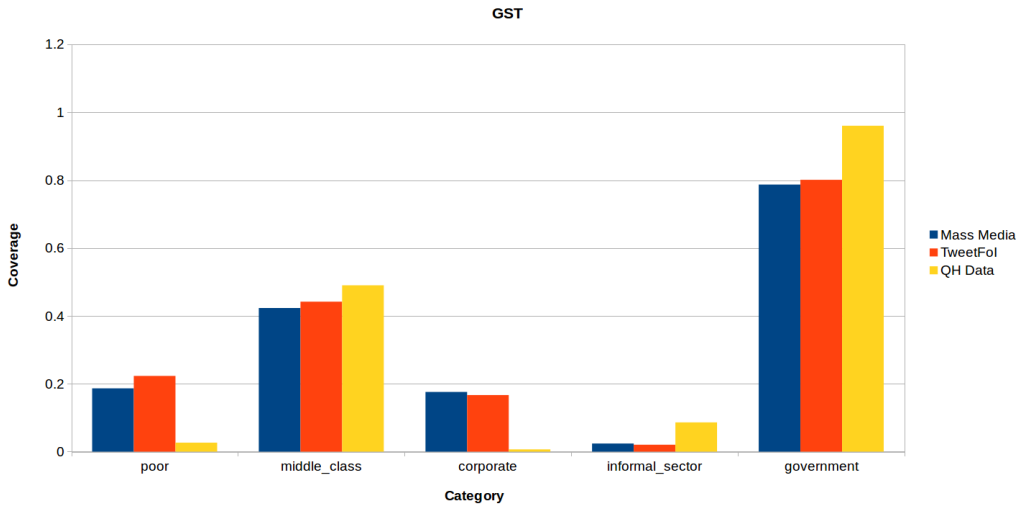


(a) Demonetisation

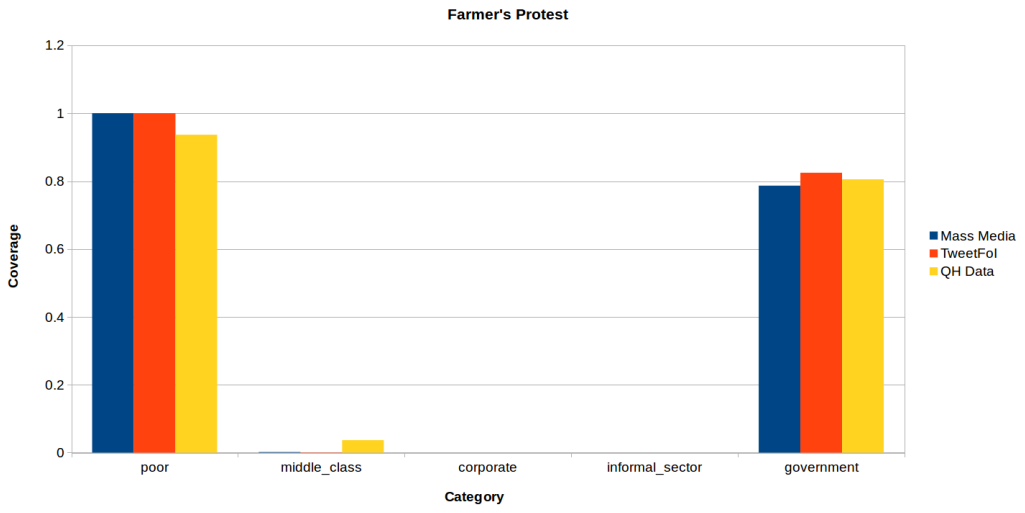


(b) Aadhaar

Figure 5.2 [RQ3] Relative constituency coverage of each policy by mass media, social media, and QH data.



(c) GST



(d) Farmers' Protest

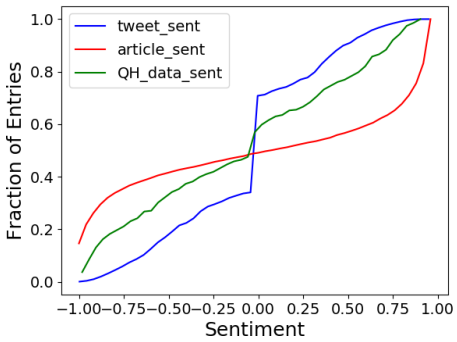
5.2.2 (RQ4:)What are the ideological slants of mass media,social media, and QH data for these policies?

In this research question,we see the ideological differences between mass media,tweets of the followers and QH data.We build three vector for this:-

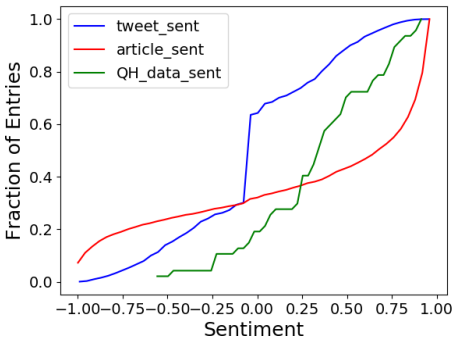
- **MM_Vec:**This contains the sentiment values of all the articles published for the policy.
- **SM_Vec:**This contains the sentiment values of all the tweets tweeted by the followers for the policy.
- **QH_Vec:**This contains the sentiment values of all the Questions asked in the Parliament in reference to a policy.

We then plot a CDF plot on **MM_Vec,SM_Vec,QH_Vec** distribution.It can be seen in Figure 5.3.We find out that the sentiments expressed by mass media are pretty polar,i.e either the articles published are too negative or too positive for a policy.On the other hand,we see that there is a steep rise in the curve at 0 value for the tweets which means that most of the tweets tweeted for a policy are neutral and do not display any biases.For QH data,we see that there is an even distribution of sentiments,i.e questions are asked in the Parliament regarding positive as well as negative aspects of a policy.

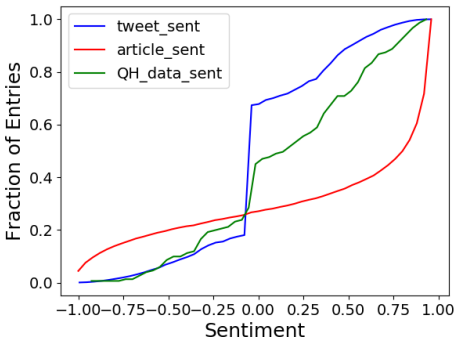
Figure 5.3 [RQ4] Sentiment Distribution of Mass Media,Social Media and QH data.



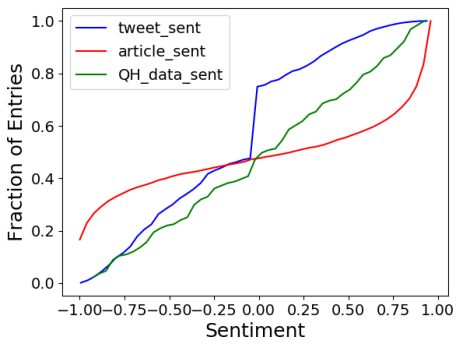
(a) Demonetisation



(b) Aadhaar



(c) GST



(d) Farmers' Protest

Chapter 6

An Attempt at Using Mass Media Data to Analyze the Political Economy Around Some Key ICTD Policies in India

6.1 Data:

We have attempted to analyse the key ICTD policies of India which are *Aadhaar,Digital India,E-Governance and Cashless Economy.*The detailed algorithm of eventwise article collection have been explained in Chapter 3.We collected **12809** articles for **Aadhaar**,**12412** articles for **Digital India**,**4853** articles for **E-Governance** and **5728** articles for **Cashless Economy**.The news articles corpus are collected since 2011.We extract entities of different types using Open Calais tool.There are entities of type *Company,Orgainisation, Location,etc* but we are concerned about the entities of type *Person* in our analysis.Next using *Elastic Search*,we obtain the top ten matching entities for every entity extracted.It uses standard string matching,aliases and context information.Next,using *jarowinkler* tool we find out the best matching entities corresponding to the main entity being considered.If the score returned is greater than or equal to 0.9,we consider the match to be the best match.After that,we further extract the sets of context information for the main entity and the best matched entity and try to find an intersection between the two sets.If the intersection is greater than a pre-defined threshold we merge those two entities along with their context information and aliases.

In this chapter,we try to analyse the coverage given by top entities for different policies and the sentiment slant of the statements made by the entities.We compute the relative coverage given by the entities using the following for-

mula:

$$relative\_coverage\_by\_the\_entity(E\_i) = \frac{count(E\_i)}{\sum_{e \in E} count(e)}$$

where count(E_i) refers to the number of *by* statements of entity(E_i).We consider the *by* statements of *top 20* covered entities.

For sentiment analysis of the statements made by the entities, we first use the Stanford CoreNLP [46] to classify the statements where the entity is mentioned in the media 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 NLP tool 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 then use the Vader tool to compute the aggregate sentiment of the *by* and *about* statement of the entities.It returns four scores-*compound,positive, negative,neutral* for a given sentence.We add up the positive sentiment scores of every *by* statement and call as TPOS and we add up the negative scores of every *by* statement and call it as TNEG.We analyse the sentiments expressed by the entities for different policies by aggregating their compound sentiment scores. *Vader* computes the compound sentiment score by adding up the positive and negative sentiment values for every word in the statement.So it may happen that the positive and the negative values may cancel out and the aggregate compound sentiment becomes almost neutral(close to zero).To capture the polarity in the statements made by the entities,we use the following formula:

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

where degpol(statement) measures how *polar* a statement made by an entity is.We calculate the degpol(entity) as the sum of the *polarity* scores of all its *by* statements.

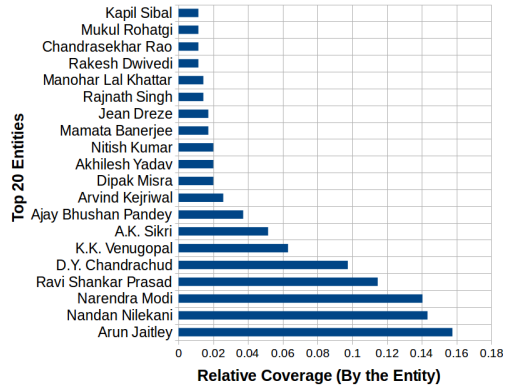
6.2 Results:

6.2.1 (RQ5:)Who are the top covered entities(politicians, academicians,activists,etc) in mass media?

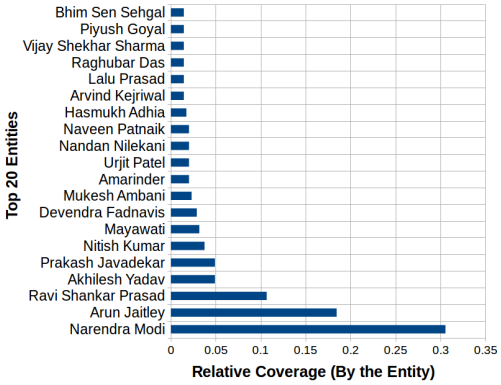
To answer this question, we calculated the relative coverages of each entity in our mass media dataset,and ranked them in decreasing order of their relative coverage.This gives us a ranked list of entities that are mentioned the most in media, corresponding to the policy issues.Figure 6.1 shows the relative coverage by top 20 entities for different policies.

We can find out some interesting conclusions from these plots.We find out that mostly the top covered(vocal) entities are politicians.So mostly political issues have been covered in the articles for all these policies.As we go deeper into the statements made,we mostly find out that there have been political blame games in their statements.Some topics like Aadhaar do see 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*, 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*.
Two evident trends were observed.a)Apart from *Politicians,Business-persons* were covered the most in mass media.b)Academicians and the social development activists were almost not covered by mass media.

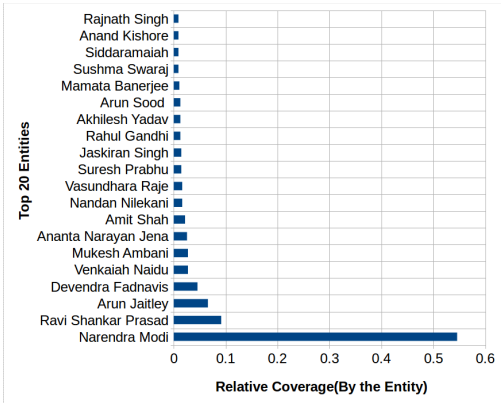
Figure 6.1 [RQ5]Plot of the relative coverage of top 20 entities for each policy for statements made by them:relative coverage is calculated as the number of statements made by the entity divided by the total number of statements by all entities,corresponding to a policy.



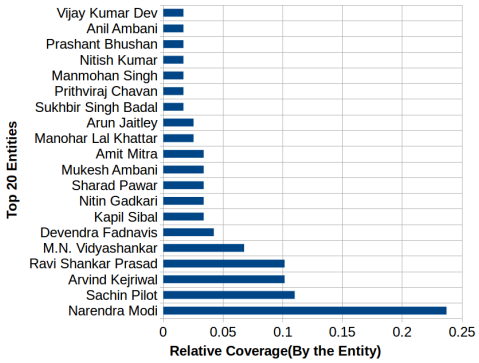
(a) Aadhaar



(b) Cashless Economy

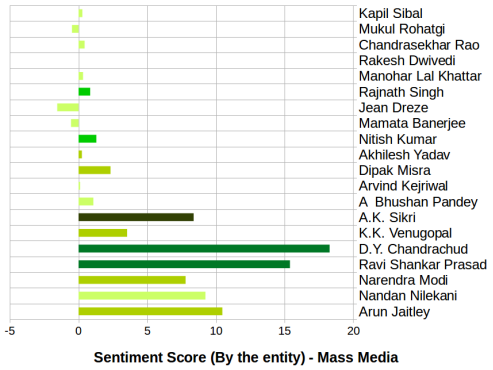


(c) Digital India

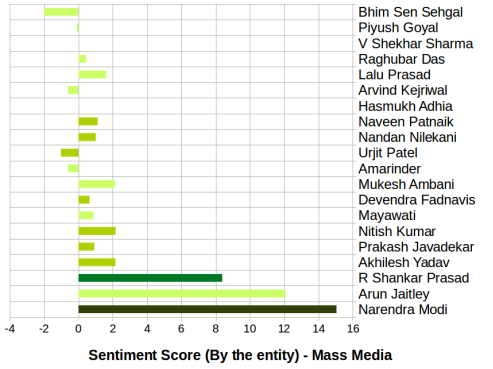


(d) E-Governance

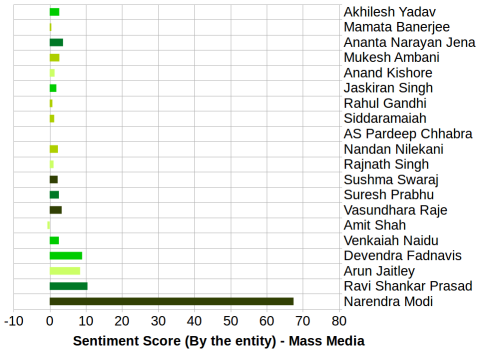
Figure 6.2 [RQ6]Plot of the aggregate sentiment, color coded on degpol for the top 20 entities with highest coverage for each policy: the aggregate sentiment/degpol is calculated as the sum total of the sentiment values/degpol values corresponding to the statements made by an entity. Higher the value of degpol (darker the color of the bar), more is the overall polarity of an entity.



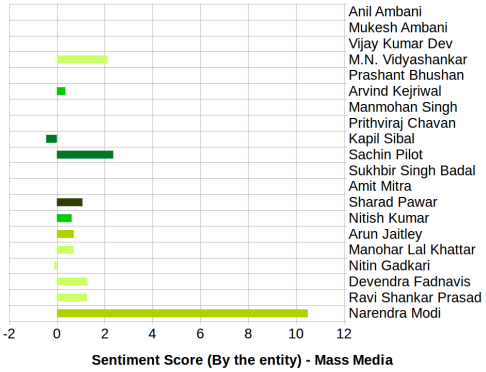
(a) Aadhaar



(b) Cashless Economy



(c) Digital India



(d) E-Governance

6.2.2 (RQ6:)What is the ideological slant of these top covered entities in mass media?

To see the ideologies 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 6.2.

In the figure, the bars on the right hand side of the zero value on x-axis rep-

resent 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 entitys sentiment.

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. 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. This is expected as Narendra Modi being the prime minister of India, was the proponent and staunch supporter of these policies. Ravi Shankar Prasad held the Electronics and Information Technology portfolio during the time we did our analysis,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.We find *Nandan Nilekani* to be less polar than *Narendra Modi*.It is pretty evident because Nandan Nilekani spoke mostly about the application and technology oriented aspects about the policy whereas Narendra Modi,being the Prime Minister engaged himself mostly with political topics.

Chapter 7

Analysis of Media Bias in Opinion Articles(Opeds) across policies

7.1 Data collection:

We analyse the media bias in opinion articles across *Aadhaar,Demonetisation, GST and Farmer's Proest.*The article extraction algorithm,aspect identification using LDA,entity and sentiment extraction algorithms are the same as explained in the preceding Chapters 4,5 and 6.The opinion articles have been collected within 2011-2018 time period.We collected **531** opinion articles for **Aadhaar**,**986** opinion articles for **Demonetisation**,**1226** opinion articles for **GST** and **558** opinion articles for **Farmer's Protest**.We tried to answer the following research questions for the opinion articles:-

- How differently are the aspects covered(overall and newspaper wise) between overall mass media and opinion articles?
- What are the ideological slants of top 20 authors for each policy in opeds?
- What is the coverage given to different entities in opinion articles and what are their sentiment slants?

7.2 Results:

7.2.1 (RQ7:)How differently are the aspects covered (overall and newspaper wise) between overall mass media and opinion articles?

In this research question,we try to see the coverage bias shown by different news sources in the opinion articles across different policies.We also find out the coverage given to different aspects across newspapers.Finally,we try to draw out some conclusions after comparing this aspect coverage with that of the overall mass media.We analysed *Aadhaar,Demonetisation and Farmer's Protest* policies.To compute relative aspect coverage per news source for a particular policy,we have used the following formula:-

$$relative\_aspect\_coverage = \frac{count(words,aspect)}{\sum_{aspect}(words,aspect)}$$

where count(words,aspect) is the total number of words of all the opinion articles belonging to the aspect published *by the news-source.* We compute the aspect coverage across news sources using the following formula:-

$$relative\_aspect\_coverage = \frac{count(words,aspect)}{\sum_{aspect}(words,aspect)}$$

where count(words,aspect) is the total number of words of all the opinion articles belonging to the aspect *across all news sources.*In the subsequent sections we write about our findings.

Aadhaar:

Findings on the Opinion articles: We find that technology and implementation driven aspects,mostly speaking about the benefits of the policy have been covered the most.In *Aadhaar*,we see that implementation and technology oriented aspects like *International Linkages/Positive Effect on Economy,Aadhaar Verification for Telecom Users/Data Leakage,Implementation of Direct Benefit Transfer Scheme* have been covered the most across news

sources and also by every news source.On the other hand we also see that aspects revealing the faults and issues with the policy like *Court Cases related to aadhaar* has been a significant coverage by every newspaper and across newspapers.We can see the visualisations per news source for Aadhaar in sectionA.8.The overall aspect coverage across newspapers can be seen in Figure 7.1.

*Comparison with the overall mass media:*In Chapter 4 and 5,we saw in details the aspect coverage given to different aspects by every news source in overall mass media(A.4),and the overall aspect coverage across news sources in overall mass media(5.1).We saw that news sources like *Hindu,DecH,TOI,Telegraph* gave a huge coverage to aspects like *Aadhaar Enrollment Centers* where mostly the articles were about the implementation issues with Aadhaar, registration problems,long queues at Aadhaar Enrollment Centers,etc.*Indian Express and Hindustan Times* gave a huge coverage to political aspects and issues like *Parliamentary debates on Aadhaar*,whereas *New Indian Express* gave a huge coverage to technological aspects related to the policy like *International Linkages,Aadhaar base verification and data leakages.*In the opinion articles,we saw the alignment to be similar across newspapers and the most coverage been given to legal issues and technology oriented aspects.We found the pattern to be similar for *New Indian Express* between opinion articles and overall mass media,whereas in other newspapers we found there were differences in the aspects they have covered the most between the opinion articles and overall mass media.

The overall aspects which have been covered the most in overall mass media have been *Parliamentary debates on Aadhaar,Court Cases related to Aadhaar,Aadhaar Enrollment Centers* where mostly legal issues related to Aadhaar like right to privacy,leakage of personal data,political issues and problems of common man have been discussed the most.On the other hand,in the opinion articles,we found that technology oriented aspects like *Positive Effect of Aadhaar on Economy*,aspects dealing with the implementation process of the policy along with legal issues have been covered the most.

Demonetisation:

*Findings on the Opinion articles:*We see that economy oriented aspect like *Negative Impact of Demonetisation on Economy* and political aspect like *Appreciation by PM for supporting Demonetisation* have been covered the most by different news sources.The trend is exactly the same when we saw the overall aspect coverage across newspapers.We can see the visualisations per news source for Demonetisation in sectionA.9.The overall aspect coverage across newspapers can be seen in Figure 7.1.

*Comparison with the overall mass media:*In Chapter 4 and 5,we saw in details the aspect coverage given to different aspects by every news source in overall mass media(A.5),and the overall aspect coverage across news sources in overall mass media(5.1).For all the news sources we saw the aspect *Appreciation by PM for supporting Demonetisation* to get covered the most.We saw *Indian Express and New Indian Express* covering legal issues like *Probes and Arrests of Black Money Hoarders* to a significant amount. *Telegraph and TOI* covered aspects more close to middle class dealing with the problems faced by common man like *Long Queues at ATMs/Banks* to a significant amount.In the opinion articles,aspects like *Negative Effect of Demonetisation and Appreciation by PM for supporting Demonetisation* were covered the most by each and every news source.

In overall mass media,political aspects like *Appreciation by PM for supporting Demon and Opposition Unites against government on Demon* were the most covered aspects across newspapers.On the other hand in the opinion articles,apart from political aspect like *Appreciation by PM for supporting Demon*,economy oriented aspects like *Negative Impact of Demon on Economy* which talked mostly about the negative impact of the policy on people and on the economical stability of the country was also given a good coverage.

Farmer's Protest:

*Findings on the opinion articles:*we see that aspects talking about the problems and concerns of the farmers like *Farmer's Distress regarding MSP for Crops,Variation in crop prices with monsoon season and Crimes/Suicides in*

Farming Community have been given the most coverage.All news sources apart from *Indian Express(pro-INC) and Telegraph and New Indian Express* have also given a significant coverage to the aspect talking about the initiatives taken by the government for farmers like *Training Programmes/Seminars to improve Farming.*We can see the visualisations per news source for Demonetisation in sectionA.10.The overall aspect coverage across newspapers can be seen in Figure 7.1.

*Comparison with the overall mass media:*In Chapter 4 and 5,we saw in details the aspect coverage given to different aspects by every news source in overall mass media(A.7),and the overall aspect coverage across news sources in overall mass media(5.1).We see *Hindu and Indian Express* covering more about aspects which deals with the initiatives taken by the government like *Loan Waiver for farmers,Training progammes to improve farming*,whereas in the opinion articles we see aspects dealing with the problems faced by the farmers getting covered the most like *Irrigation Concerns for farmers,Variation in crop prices with monsoon season* getting covered the most.We see *DecH, Hindustan times,NIE and IE* giving most coverage to aspects like *Protests by farmers,crimes and suicides by farmers* which talked about the problems faced by the farmers and the negative impact of that on the entire farming community,whereas in the opinion articles *DecH and HT* covered *Training Programmes to improve farming* the most which talked about the steps taken by the government for the farmers.In the opinion articles,we also see *IE and NIE* covering aspects similar to that in the overall mass media.We see *Telegraph* giving most coverage to aspects like *Training programees for farmers,Protests by farmers* which talked about the steps taken by the farmers against the government as well as about the initiatives taken by the government.In the opinion articles,we see *Telegraph* covering only the issues of the farmers and the steps taken by them against the government.

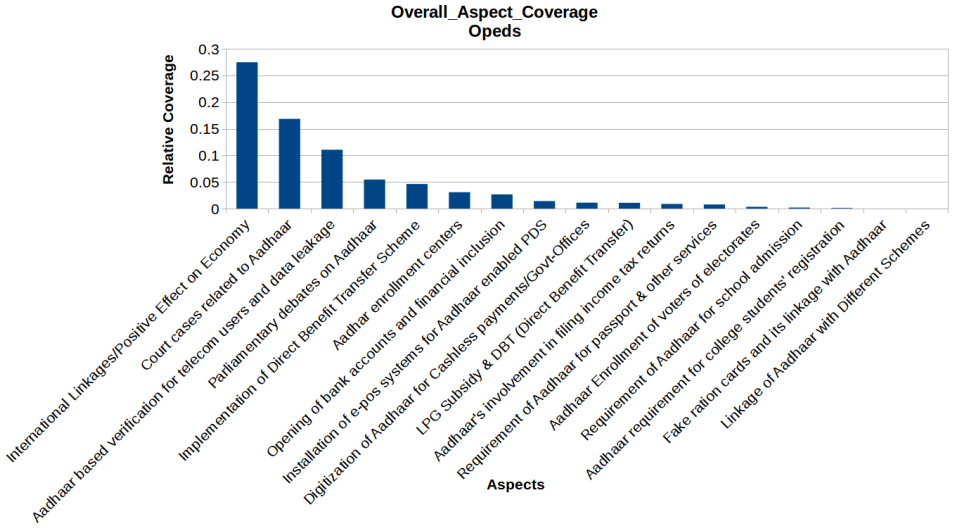
In overall mass media,the highest covered aspect across all newspapers is **Disbursement of loans and subsidies to farmers by banks**.This aspect included discussions about loans provided to farmers by banks under different government schemes.The second most covered aspect across news sources is **Irrigation concerns and water pollution affecting farming** which included discussions about the problems faced by farmers in irrigation, due

to water pollution through industrial effluents.In the opinion articles,we see that apart from the above mentioned two aspects,aspects dealing with the steps taken by the government like *Training Programmes for farmers* also got a significant coverage.

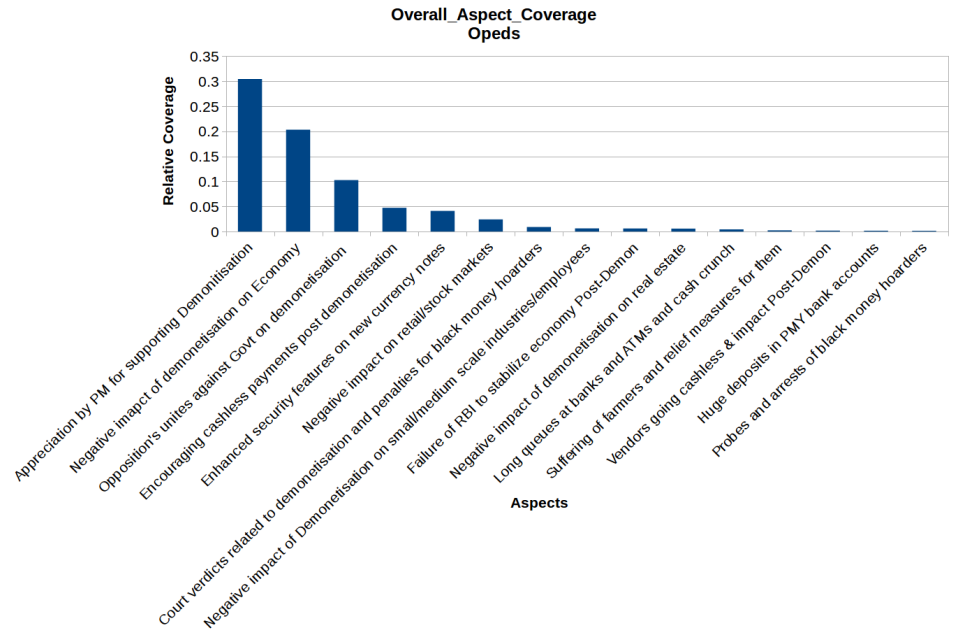
For all the policies,we see that technology driven aspects which are more about the implementation and benefits of the policy are covered the most.We see that aspects concerned about the economy of India,aspects pointing out the technological benefits and faults have been given the most coverage across all the news sources.We also see that political oriented aspects have also been an equally high coverage.We find a strong alignment in the coverage given by different news sources to different aspects of the policies being studied.Let us look into each policy one by one.

For *Aadhaar*,we see that implementation and technology oriented aspects like *International Linkages/Positive Effect on Economy,Aadhaar Verification for Telecom Users/Data Leakage,Implementation of Direct Benefit Transfer Scheme* have been covered the most across news sources.On the other hand we also see that political aspect like *Court Cases related to aadhaar* has been a significant coverage by most newspapers.For *Demonetisation*, we see that economy oriented aspect like *Negative Impact of Demonetisation on Economy* and political aspect like *Appreciation by PM for supporting Demonetisation* have been covered the most by different news sources.Next,for *Farmer's Protest*,we see that aspects talking about the problems and concerns of the farmers like *Farmer's Distress regarding MSP for Crops,Variation in crop prices with monsoon season and Crimes/Suicides in Farming Community* have been given the most coverage.All news sources apart from *Indian Express(pro-INC) and Telegraph* have also given a significant coverage to the aspect talking about the initiatives taken by the government for farmers like *Training Programmes/Seminars to improve Farming*.We can see the visualisations for aspect coverage per news sources for Aadhaar in section A.8,for Demonetisation in section A.9 and for Farmer's Protest in section A.10.

Figure 7.1 [RQ7]Aspect Coverage across news sources for opinion articles.

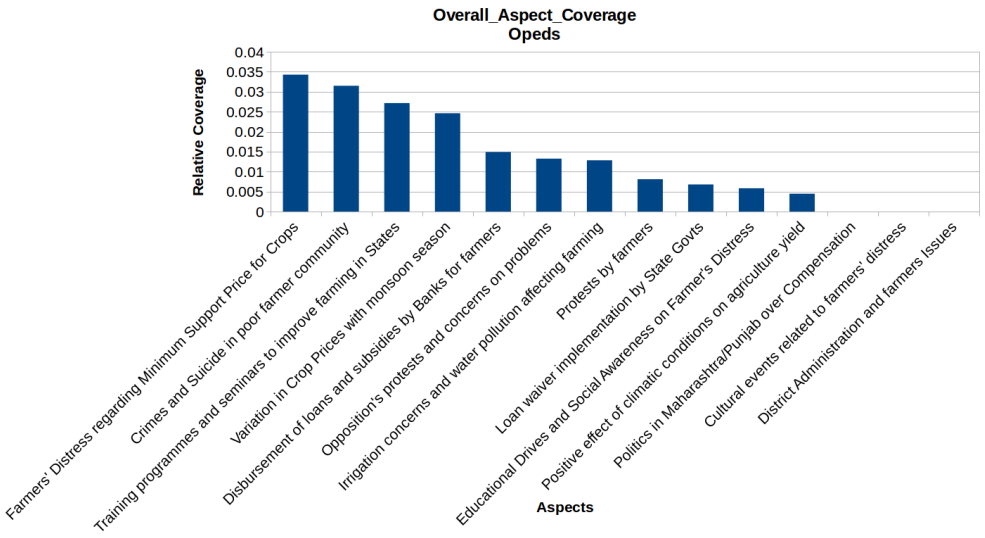


(a) Aadhaar



(b) Demonetisation

Figure 7.1 [RQ7]Aspect Coverage across news sources for opinion articles.



(c) Farmer's Protest

7.2.2 (RQ8:)What are the ideological slants of top 20 authors for each policy in opeds?

In this research question,we find the most frequent authors writing the opinion articles and try to see their ideological slants.We compute the aggregate sentiment of an author by adding up the compound sentiment scores returned by Vader.To measure the polarity of the articles written by the authors,we use the following formula:-

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

where degpol(article) refers to the polarity score of the article.In figure 7.2,we see the most frequent authors writing opinion articles for each policy.For all the policies,we see the chief editors of various newspapers,academicians and journalists to be the most frequent authors of the opinion articles.In figure 7.3,we see the sentiment slant of these authors.For *Aadhaar*,we see several famous editors from different newspapers writing opeds on Aadhaar issue.In

Aadhaar we see Times Of India's Senior Assistant Editor,*Sanjiv Shankaran* displaying negative sentiment towards this policy.Mostly he spoke about the implementation issues and the inability of the governmental infrastructure to implement the utilities of this policy.In one article he wrote that *Union governments recent notification on some Aadhaar-related regulations is unlikely to make the 12-digit identifier a mandatory requirement for all public welfare schemes. Inadequate supporting infrastructure will ensure that Aadhaar will be used in conjunction with other forms of identity in large parts of India.The notification acknowledges the problems. Hence, the caveat that government agencies which require people to produce Aadhaar need to arrange for it in case a beneficiary does not have one. This covers the people who do not have an Aadhaar number. But other problems which showed up in pilot studies suggest that we have a long way to go before Aadhaar can fulfill its potential.For Aadhaar to work without frictions, supporting infrastructure in the form of point-of-sales machines and internet connectivity need to be excellent. So far, they have been a letdown in some parts of India.*Next,we see *Jean Dreze* showing a negative sentiment slant for Aadhaar,which is natural because Jean Dreze has been one person who has criticized Aadhaar strongly.We see *Nandan Nilekani,former UIDAI chief* speaking positively for Aadhaar.He spoke mostly about the technological aspects and how can this policy benefit common people.In one of the articles he wrote,he said,*Lok Sabhas passing of the Aadhaar Bill lays the foundation for an efficient, effective and modern welfare state. The reform of Indias subsidy regime will now be taken beyond LPG to include kerosene, fertiliser, food, water, electricity, interest subvention, cash support to farmers instead of the distortionary minimum support prices and much else.Aadhaar will not only deliver minimum consumption levels to those who need it but make resources available from the subsidy saved from misdirection and fraud. However, there is another equally large and immediate opportunity that has now presented itself thanks to the future of Aadhaar being unequivocally settled; the use of the Aadhaar platform to revive credit and kick-start the Indian economy.*

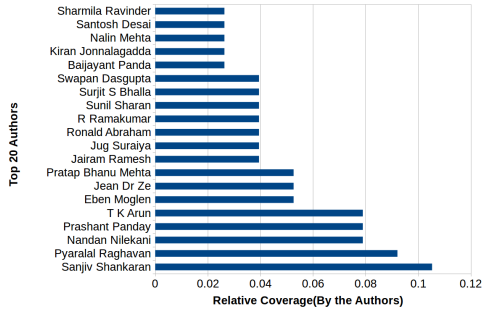
For *Demonetisation*,we saw many politicians,editors and journalists writing the opinion articles frequently.*Kanishka Singh*,who belongs to INC is a political aide to *Rahul Gandhi*,spoke negatively about this policy.He mostly spoke

about the hardships and the difficulties which was being faced by the common people and that Demonetisation will only be able to detect all black money in cash and not in any other forms.In one of his articles he wrote,*Top experts in the country estimate that black money in cash is only about 1/20th of the black economy.Demonetisation has become the demon word for much of India's population in recent weeks. Though many agree with the policy, others complain about the manner it was implemented. It has been tough, confusing, desperate and smeared with dollops of patriotism, perhaps, to deviate the minds from its intricacies. The question, however, is: will demonetisation actually put an end to black money, which itself is a complicated entity and not restricted to currency. At best, the move is only curbing the black wealth accumulated in the form of cash and leaving out a massive amount of black wealth converted in form of gold, shares, foreign currency and other financial instruments away from the effects of demonetisation.*People like *Sagarika Ghose,consulting editor of TOI,P.Chidambaram,INC politician*, all wrote frequently and negatively about the policy focusing on the fact that Demonetisation will only increase the difficulties of common people and will not be able to curb black money.We also saw that *Sa Aiyar,journalist and consulting editor at the Economic Times* speaking positively about the policy.He wrote articles of mixed sentiments.Stating the problems that were faced by citizens,he also repeatedly stated that black money in foreign banks will eventually return to country and to our government.

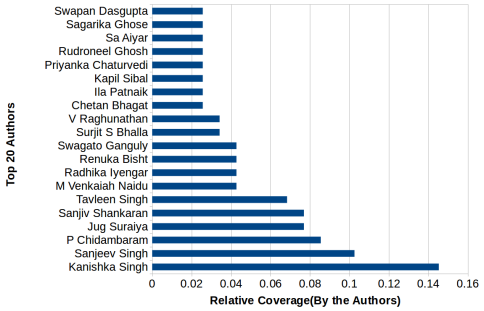
For *Farmer's Protest*,we see the authors speaking negatively about this issue.Among the most frequent authors,we see many journalists and academicians.*Ashok Gulati Siraj Hussain,an agricultural economist*,has spoken highly negatively about this issue.He has spoken mostly about the plight of the farmers post demonetisation.He wrote that the cash crunch post demonetisation has mostly hit the unorganised sector of the society which deals with liquid cash everyday.Cash crunch hadn't allowed farmers to purchase required inputs for their crops which has caused hardships to farmers.He wrote,*Demonetisation has evoked sharp and extreme reactions from various quarters. At one end, we have nation-wide surveys of C-Voter and the PMs App, showing 88-90 per cent people supporting it, and at the other, we have strong protests in and outside Parliament by several Opposition parties. Man-*

*mohan Singh, former prime minister, went even to the extent of dubbing it as a case of monumental mismanagement and organised loot and plunder, which may cost the economy 2 percentage points of GDP growth.One may differ from both extremes, but almost all agree that in the immediate run, the unorganised sector has been hit hard as it runs largely on cash. And within the unorganised sector, farmers are suffering as they do not have enough cash to purchase the required inputs for the sowing of rabi crops.The rabi area, as on November 25, was about 7 per cent lower than the average of the last five years, and fertiliser consumption during November 1-28 is about 12 per cent lower than the average of the last three years. Comparison with last year is not appropriate as it was a severe drought year. Although sowing is still on and it may soon catch up to normal levels, it is true that the cash crunch is causing quite a bit of inconvenience and hardship to farmers. The government needs to act on a war footing to increase liquidity and promote e-payments in rural areas.*We also *Sa Aiyar,famous journalist* showing negative sentiments about this issue,but the articles he wrote were not about hurling acquisitions at the government and speaking about the problems of poor,but were mainly targeted at the Congress Party and how the party was using this farmer's issue against the government and to increase their vote banks.In one of the articles,he wrote,*The vast majority of suicides are of non farmers. Why are their deaths treated as lesser tragedies than those of farmers?Because presenting farm suicides as a single mass tragedy can win awards for journalists, TRPs for TV anchors, donations for NGOs opposing commercial crops and globalization, slogans for leftists attributing everything to class war, and votes for opposition parties. Many states now compensate suicide-hit families, delighting moneylenders who had lent to these families and can now use muscle to claw back their dues from the compensation money.*

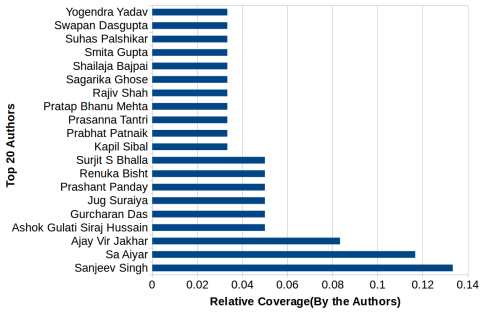
Figure 7.2 [RQ8]Plot of the relative coverage of top 20 authors for each policy for opinion articles written by them:relative coverage is calculated as the number of articles written by the author divided by the total number of written by all authors,corresponding to a policy.



(a) Aadhaar

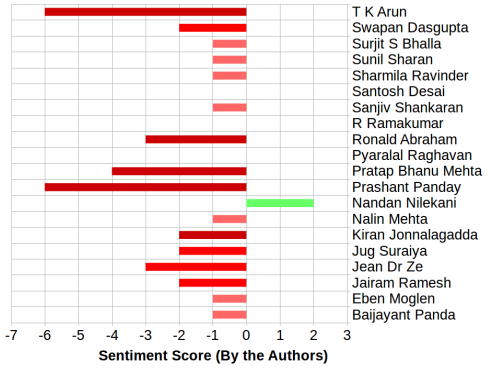


(b) Demonetisation

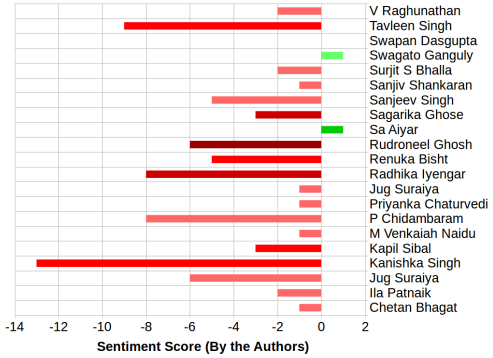


(c) Farmer's Protest

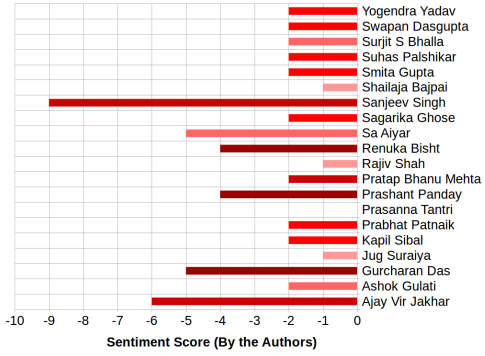
Figure 7.3 [RQ8]Plot of the aggregate sentiment, color coded on degpol for the most frequent 20 authors with highest number of opeds written for each policy: the aggregate sentiment/degpol is calculated as the sum total of the sentiment values/degpol values corresponding to the articles written by an author. Higher the value of degpol (darker the color of the bar), more is the overall polarity of an author.Shades of Red colour indicate negative aggregate sentiment and shades of Green indicate positive aggregate sentiment



(a) Aadhaar



(b) Demonetisation



(c) Farmer's Protest

7.2.3 (RQ9:)What is the coverage given to different entities in opinion articles and what are their sentiment slants?

In this research question,we try to find out which entities have been covered in the opinion articles and what were the sentiment slants in the statements made by them.To extract entities and to find out their *by and about* statements,we have followed the same process as have been explained in Chapter 6.Computing relative coverage and their aggregate sentiments and polarity have also been explained in Chapter 6.In figure 7.4,we can see the top 20 covered entities in the opinion articles.In figure 7.5,we can see the sentiments expressed by them.We see that mostly politicians are covered,but there is some coverage given to people from judiciary and activists.

In *Aadhaar*, we see politicians like *Narendra Modi,Arun Jaitley,Sonia Gandhi, Piyush Goyal* getting covered the most along *Mukul Rohatgi,senior advocate and former attorney general of India.*We see positive sentiment slants in all of their *by* statements.In one of his *by* statements,Narendra Modi said,*i am sure most of you are already aware of aadhaar which is the unique biometric identity of a person through better targeting of subsidies the jam (jan dhan-aadhaar-mobile) trinity has prevented leakages to the tune of nearly 10 billion dollars so far.*Arun Jaitley said,*the aadhaar bill claimed that a sum of rs 14672 crore was saved in lpg subsidy for the year 2014-15 under the direct benefits transfer scheme.*Mukul Rohatgi told the supreme court during a hearing on the new uses of the aadhaar card that *citizens cannot claim absolute right over their bodies.*In all of their statements,we see that all the entities spoke abut the benefits of aadhaar as a policy and about the technology oriented side of this policy.

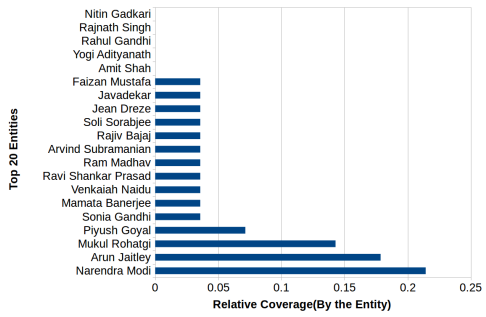
Next,in *Demonetisation*, we see that mostly politicians like *Narendra Modi, Rahul Gandhi,Arun Jaitley* have been covered with the only exception being *Urjit Patel,former RBI Governor.*We see *Rahul Gandhi,Congress President* expressing a positive sentiment slant but that was mostly because he had been sarcastic in his statements and our sentiment tool hasn't been able to pick up the real intent.We see *Nitish Kumar* expressing a negative slant,but in his statements he has strongly supported the move and has condemned

the opposition for their incessant protests.We see *Urjit Patel* expressing a negative sentiment slant and that was because he refused to agree to government's plea of changing the financial structure of RBI like keeping more liquid cash,as he said that this would lead to financial instability and he would not allow that to happen.He refused to meet and negotiate with government's officials.In one his statements he said,*members of the monetary policy committee had refused to meet finance ministry officials for a policy review meeting.*

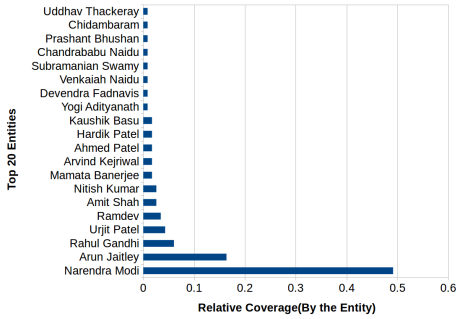
In *GST*,we see similar trends of politicians getting the most coverage.*Arun Jaitley,then Finance Minister* had expressed strong positive sentiments towards the policy and had mainly made statements related to implementation oriented issues of the policy.In one of his statements,he said,*only one difficult issue remains pending to enable the implementation of the new indirect tax regime.*In another statement,he said,*the union cabinet will soon take up the four laws that the centre has to steer through parliament, while the respective state governments will take up the state gst law.*On the other hand,*Sachin Pilot,Congress Politician* had expressed strong negative sentiments against this policy.In one his statements,he said,*people have defeated bjp because of the 'gag' ordinance,leasing of mines, caste violence and the failure to waive farm loans.*

In *Farmer's Protest*,politicians like *Devendra Fadnavis,Narendra Modi* were covered the most.*Devendra Fadnavis* in his statements have been seen very concerned and supportive regarding the plight of the farmers.In one of his statements he said,*he respected them and had no problem in accepting most of their demands, leading the farmers to end the stir and board special trains to go back home.*Another statement read that,*less than 48 hours after the strike started, fadnavis announced a rs 30,000 crore loan waiver the biggest ever in maharashtra.*

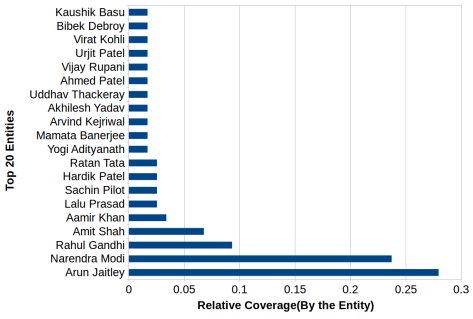
Figure 7.4 [RQ9]Plot of the relative coverage of top 20 entities for each policy for statements made by them in the opinion articles:relative coverage is calculated as the number of statements made by the entity divided by the total number of statements by all entities,corresponding to a policy.



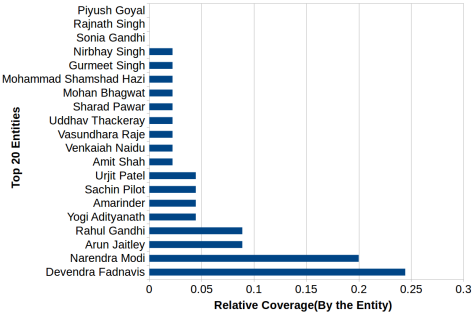
(a) Aadhaar



(b) Demonetisation

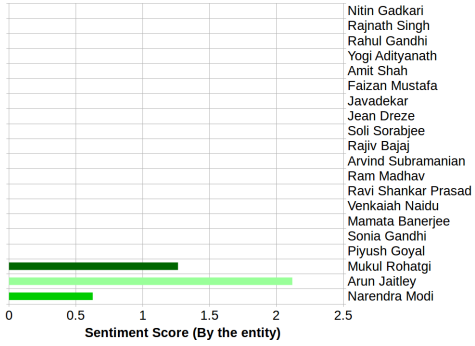


(c) GST

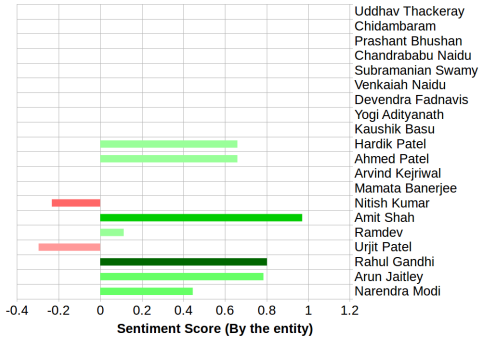


(d) Farmer's Protest

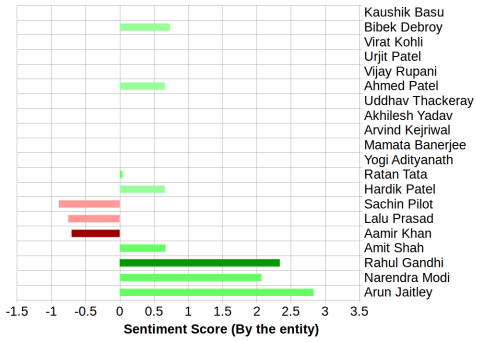
Figure 7.5 [RQ9]Plot of the aggregate sentiment, color coded on degpol for the top 20 entities with highest coverage for each policy in the opinion articles: the aggregate sentiment/degpol is calculated as the sum total of the sentiment values/degpol values corresponding to the statements made by an entity. Higher the value of degpol (darker the color of the bar), more is the overall polarity of an entity.Shades of Red colour shows negative aggregate sentiment and shades of Green color shows positive aggregate sentiment



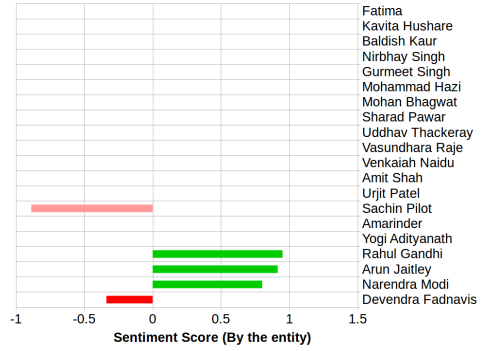
(a) Aadhaar



(b) Demonetisation



(c) GST



(d) Farmer's Protest

Chapter 8

Conclusion

In our work,we have analysed media bias with respect to different aspects.In our first research question,we analysed the coverage bias given by different news sources to different aspects.We observed that,there is a strong similarity between the coverage given to different news sources in mass media.we find that only Indian Express(IE) differs the most from the mean aspect coverage given by all news sources.In Farmer's Protest,we see that the newspapers differ the most in terms of coverage given to different policy aspects with some media houses giving more coverage to political aspects whereas some giving more coverage to implementation oriented aspects.

Next,we studied the ideological alignment between the news sources and their audience in twitter media.We found out that the newspapers are aligned with their audiences.We found a steep rise at zero value in the CDF plot for sentiment difference between these two media.

Next,we analysed the coverage given to different aspects and constituencies across mass media,social media and Question Hour Data.We see that in terms of coverage given to constituencies,poor people have been ignored the most with most coverage given to middle class and government.On all the policies, the QH focuses on the mechanics and procedural aspects related to policies. In demonetization, the focus is on procedures to restore normalcy and in cross-checking the narrative behind the policy. In Aadhaar, the focus again is on the operations, with a greater tilt towards the middle class. In GST,the focus is entirely on the rates and processes of relevance to the traders, and not the consumers. In Farmers Protests, the focus is again on the mechanics of loans and loan waivers, and not on the structural issues in agriculture. The ruling party tends to focus more on procedural aspects, while the opposition tends to focus on the problems in these procedures. However, the focus of the policy makers remains on procedures w.r.t. each policy.From our analysis of Aadhaar and Farmers Protests therefore,we find that the parliament is not getting to the depths of the issues relevant to the poor. In Aadhaar, even

the media has not reached this depth. In Farmers Protests, the mass media and the social media both pay more attention to the structural problems in agriculture. However, this trend does not make its way prominently to the QH. Similarly in GST, there is inadequate attention paid to consumers in general. The focus of the parliament seems to have been entirely driven by increasing tax revenues and to bring about greater formalization, without thinking much about the consumers.The mass media too did not focus much on consumers, or on the informal sector, which has been deeply impacted since its space is being taken over by increasing formalization of the industries. demonetisation did see more attention to the issues of the poor and the middle class, but the coverage in the media was highly politicized around the narrative of demonetisation, and this was also carried forward in the QH.By and large therefore, there seems to be an alignment between the mass media, social media, and the parliament in terms of the constituencies on which they focus the mass media and social media give more coverage to stories and experiences of the key constituencies, while the QH focuses on the mechanics and procedural aspects relevant to these constituencies. Which constituencies to focus on, however, seems to be driven by different agendas. demonetisation was political, and any associations to constituencies was made with a view of legitimizing or de-legitimizing the policy. GST was fiscal, and mostly covered the corporate constituency. Aadhaar and farmer protests were more about the people, but less about the poor (Aadhaar) or attention to deep understanding of the issues of the poor (farmer protests). The social media also does not offset this in any significant manner, which indicates the lack of interest or understanding even by the educated online population of issues that directly concern them. Except for Farmers Protests, the social media also does not show any interest or understanding of issues of the poor who are not online. When the different constituencies are indeed taken into account in the mass media and in the parliament, it is in fact in a politicized manner (demonetization). When issues are not politicized however, like with the other policies, our data shows that democratic operations are far from ideal. This does not bode well for a democratic country like India, because it shows that the representation is broken both in the media as well as in the parliament.

Next,we analyse the sentiment distribution across these mass media,social media and QH data.We find out that the articles are highly polar whereas the tweets are mostly neutral.The sentiments expressed by Parliamentary questions are less polar.

Next,we also analyse the coverage given to different entities in mass media and their ideological slants.We see a huge coverage been given to politicians and business-persons and very less coverage given to academicians and activists across policies.

Lastly,we analyse the media bias and entity coverage on the opinion articles written in mass media.We see that mostly technical,economical and procedural aspects have been covered across policies.We also analyse the sentiments of the most frequent authors writing the opinion articles.Politicians are given less coverage in the opinion articles and more coverage have been given to famous editors of newspapers,journalists and people from judiciary.

Appendix A

Media Analysis

A.1 Articles extracted per news-source for each event

News-source	Demonetisation	Aadhaar	GST	Farmers' Protest
Hindu	2905	4550	3530	5596
HT	1995	833	1312	1359
TOI	4719	3713	4394	3379
TeleG	1313	781	1111	376
NIE	2677	820	2356	731
IE	2758	800	1044	1255
DecH	1398	1129	1726	1019

Table A.1: Count of articles extracted for each news-source for events: Demonetisation, Aadhaar, GST, and Farmers' Protest

A.2 Average SentiStrength per news-source for each event

News-source	Demonetisation	Aadhaar	GST	Farmer's Protest
HINDU	-0.750	-0.122	-0.509	-0.887
HT	-0.802	-0.555	-0.653	-1.147
TOI	-0.784	-0.416	-0.523	-1.065
TeleG	-0.765	-0.590	-0.635	-1.141
NIE	-0.763	-0.367	-0.447	-1.221
IE	-0.891	-0.480	-0.593	-1.093
DecH	-0.740	-0.341	-0.417	-1.095

Table A.2: Average SentiStrength of each news-source for four events

A.3 (Aspect,Category) mapping of all Events

Aspect	Poor	Middle Class	Corporate	Informal Sector	Govt
Training Programmes for Creating GST Awareness	0	0	1	0	0
Positive Effect of GST on GDP and Economy	0	0	0	1	1
Uncertainty in Impact of GST on consumer Goods	0	1	0	0	0
Revolts/Confusion with GST Rate Slabs	0	0	-1	0	1
GST Bill implementation in State Assemblies	0	0	0	0	-1
Discussion in Parliament in support of GST	0	0	0	0	1
Problems with working capitals faced by traders, and its impact on imports/exports	0	0	-1	0	1
Traders hassles to meet GST registration deadline	0	0	-1	0	1
Sensex/Market rejoices on GST roll out	0	0	1	0	0
Non conclusive discussion b/w center and state regarding GST	0	0	0	0	1
Nationwide Protest to include Petroleum products under GST	-1	-1	0	0	1

Table A.3: (Aspect,Category) mapping for *GST*: A cell value of 1 represents a positive alignment, -1 represents a negative alignment, and 0 represents no alignment

Aspect	Poor	Middle Class	Corporate	Informal Sector	Govt
Requirement of Aadhaar for passport and other services	0	1	0	0	1
Fake ration cards and its linkage with Aadhaar	1	1	0	0	1
Installation of e-pos systems for Aadhaar enabled PDS	-1	-1	0	0	1
Digitization of Aadhaar for Cashless payments/Govt-Offices	0	1	0	0	1
Requirement of Aadhaar for school admission	0	1	0	0	1
Linkage of Aadhaar with different schemes	0	1	0	0	1
Implementation of Direct Benefit Transfer Scheme	1	1	0	0	1
Aadhaar based verification for telecom users	0	1	0	0	-1
Opening of bank accounts and financial inclusion	1	1	0	0	1
LPG Subsidy and DBT (Direct Benefit Transfer)	1	1	0	0	1
Aadhaar Enrollment of voters of electorates	0	0	0	0	1
Parliamentary debates on Aadhaar	0	0	0	0	-1
Aadhaar's involvement in filing ITR's	0	1	0	0	1
International Linkages/Positive Effect on Economy	0	0	1	0	1
Court cases related to Aadhaar	-1	-1	0	0	1
Aadhaar requirement for college students registration	0	1	0	0	0
Aadhar enrollment centers and UIDAI operations	0	1	0	0	1

Table A.4: (Aspect,Category) mapping for *Aadhaar*: A cell value of 1 represents a positive alignment, -1 represents a negative alignment, and 0 represents no alignment

Aspect	Poor	Middle Class	Corporate	Informal Sector	Govt
Failure of RBI to stabilize economy Post Demonetisation	0	0	0	0	1
Negative impact on Small/Medium scale industries and employees	-1	-1	0	-1	1
Long queues at banks and ATMs and cash crunch	-1	-1	0	0	1
Court verdicts and Penalties for black money hoarders	0	-1	0	0	0
Negative impact on Retail/Stock markets	0	0	1	0	0
Vendors going cashless and impact Post Demonetisation	-1	-1	0	-1	0
Probes and Arrests of black money hoarders	0	0	0	0	-1
Suffering of farmers and relief measures for them	-1	0	0	0	1
Huge deposits in PMY bank accounts	1	1	0	0	1
Negative impact of demonetisation on Economy	-1	-1	-1	-1	0
Appreciation by PM for supporting Demonetisation	1	1	0	0	1
Negative impact of demonetisation on Real estate	0	-1	0	0	0
Enhanced security features on new currency notes	0	0	0	0	0
Opposition unites against government on Demonetisation	0	0	0	0	-1
Encouraging cashless payments post Demonetisation	-1	-1	0	0	1

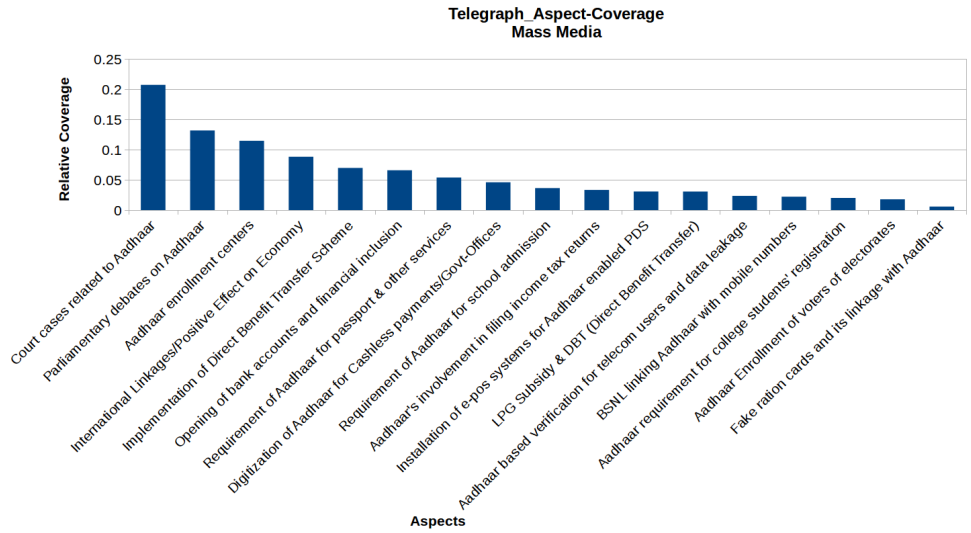
Table A.5: (Aspect,Category) mapping for *Demonetisation*: A cell value of 1 represents a positive alignment, -1 represents a negative alignment, and 0 represents no alignment

Aspect	Poor	Middle Class	Corporate	Informal Sector	Govt
Positive effect of climatic conditions on agriculture yield	1	0	0	0	0
Opposition's protests and concerns on problems	-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	1	0	0	0	-1
Variation in Crop Prices with monsoon season	-1	-1	0	0	0
Cultural events related to farmers' distress	1	0	0	0	0
Disbursement of loans and subsidies by Banks for farmers	-1	0	0	0	1
Crimes and Suicide in poor farmer community	-1	0	0	0	1
Protests by farmers	-1	0	0	0	1
Loan waiver implementation by State Govts	1	0	0	0	1
Training programs and seminars to improve farming in States	1	0	0	0	0
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

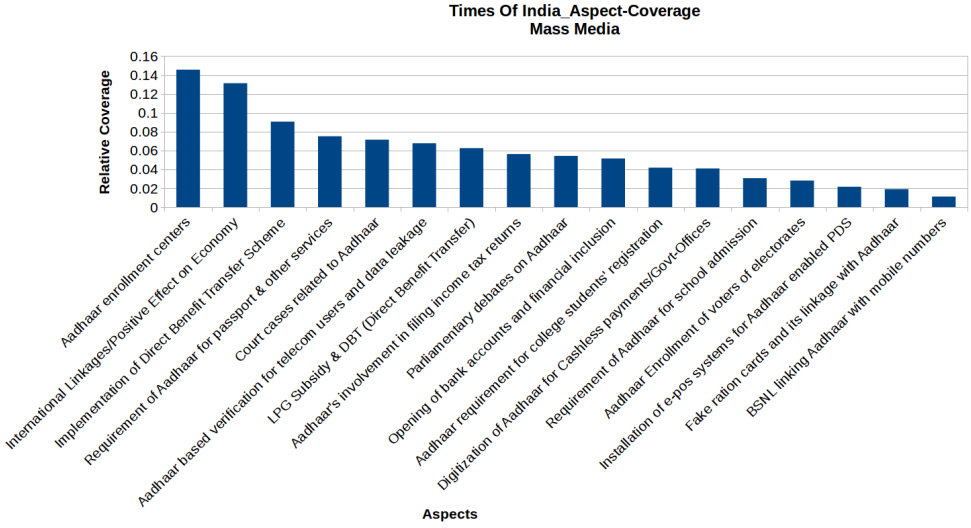
Table A.6: (Aspect,Category) mapping for *Farmer's Protest*: A cell value of 1 represents a positive alignment, -1 represents a negative alignment, and 0 represents no alignment

A.4 Aspect Coverage for Aadhaar:

Figure A.1 [RQ1] Newspaper wise Relative Aspect coverage for Aadhaar in mass media.

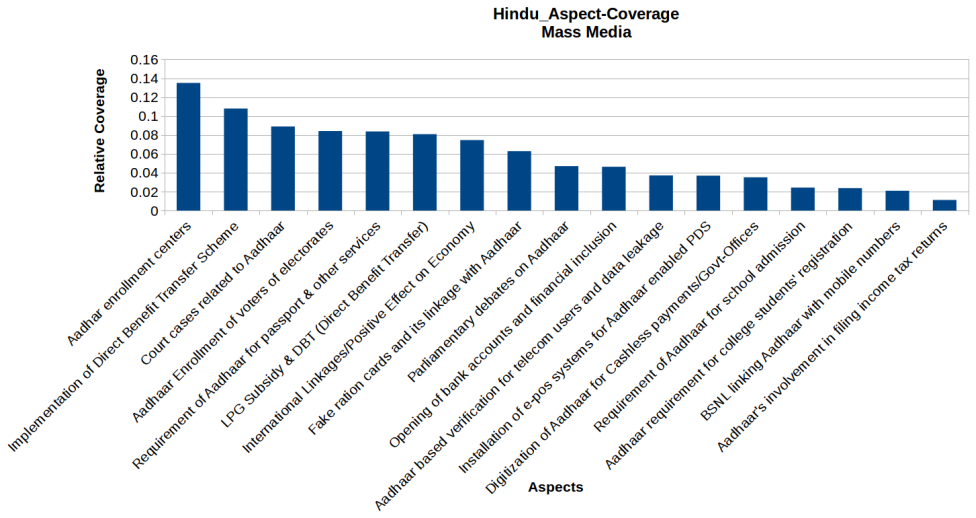


(a) TeleG

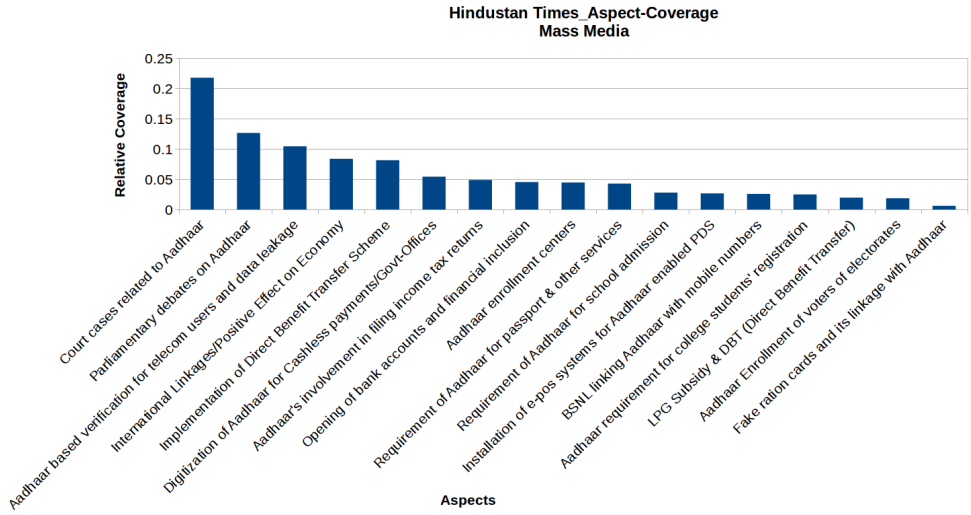


(b) TOI

Figure A.2 [RQ1] Newspaper wise Relative Aspect coverage for Aadhaar in mass media.

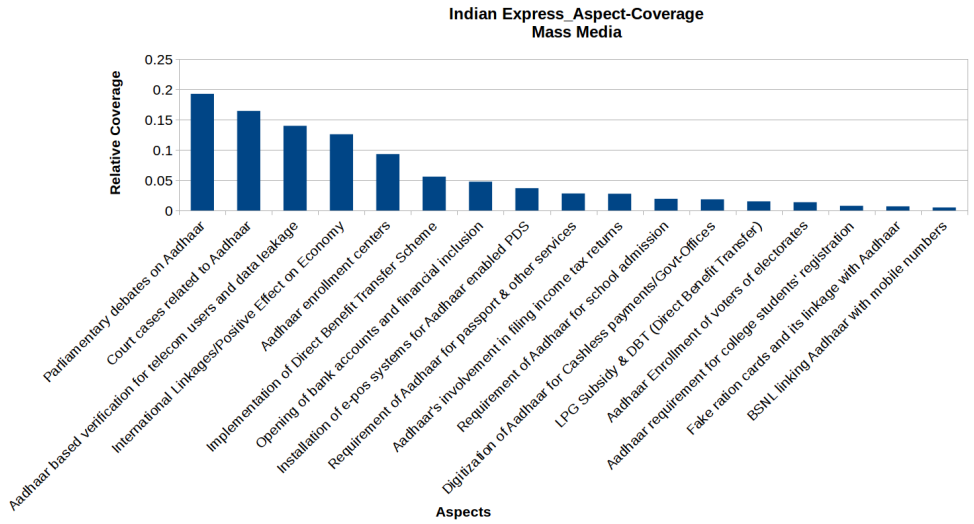


(a) The Hindu

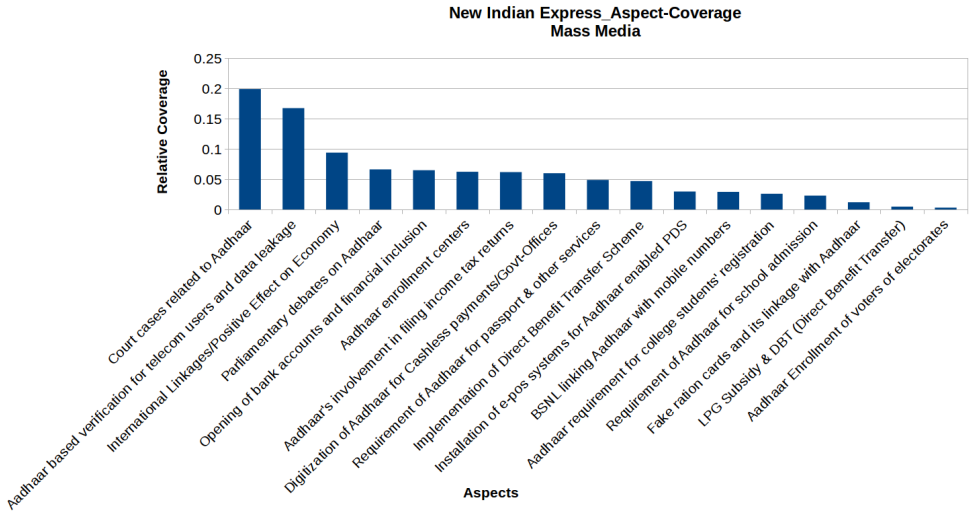


(b) HT

Figure A.3 [RQ1] Newspaper wise Relative Aspect coverage for Aadhaar in mass media.

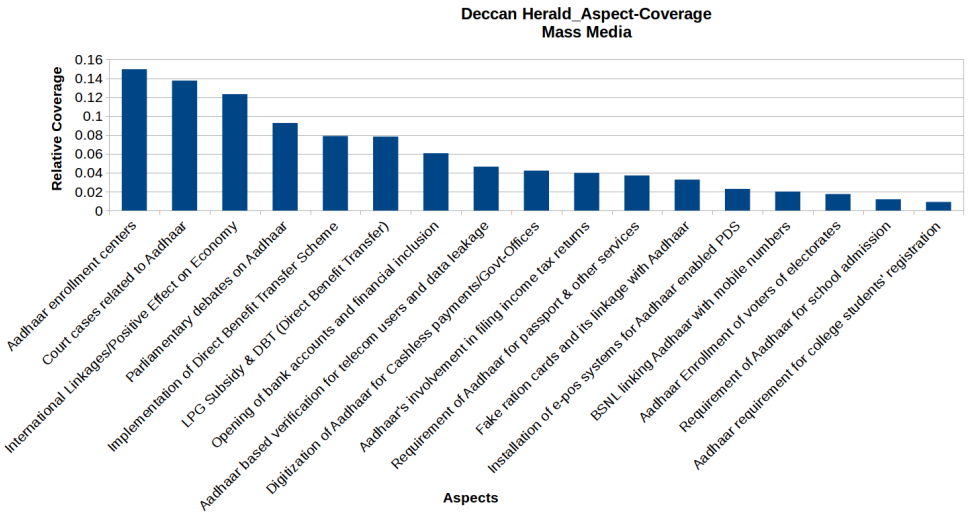


(a) IE



(b) NIE

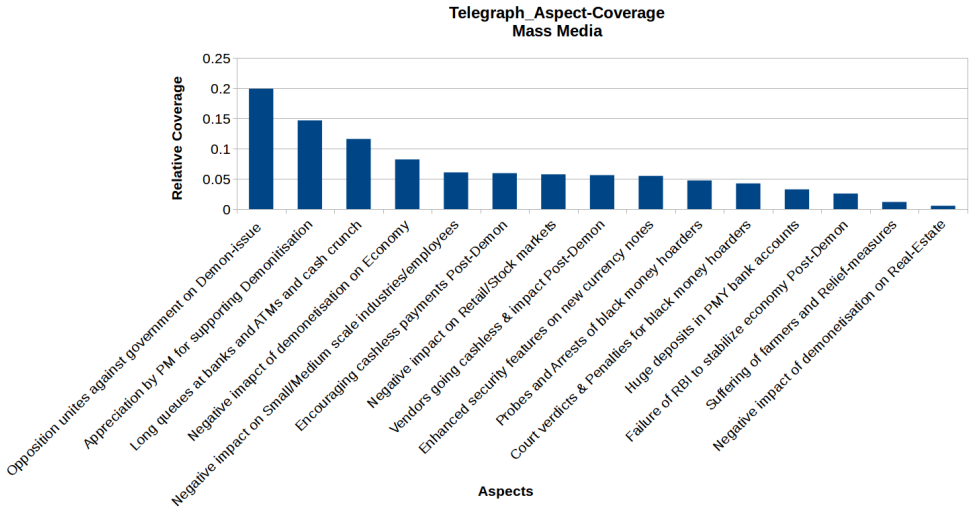
Figure A.4 [RQ1] Newspaper wise Relative Aspect coverage for Aadhaar in mass media.



(a) DecH

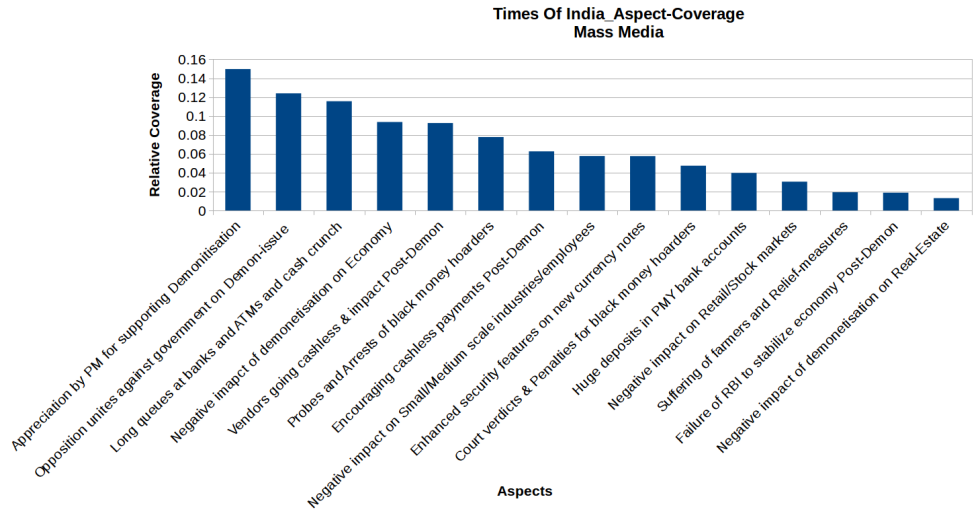
A.5 Aspect Coverage for Demonetisation:

Figure A.5 [RQ1] Newspaper wise Relative Aspect coverage for Demonetisation in mass media.

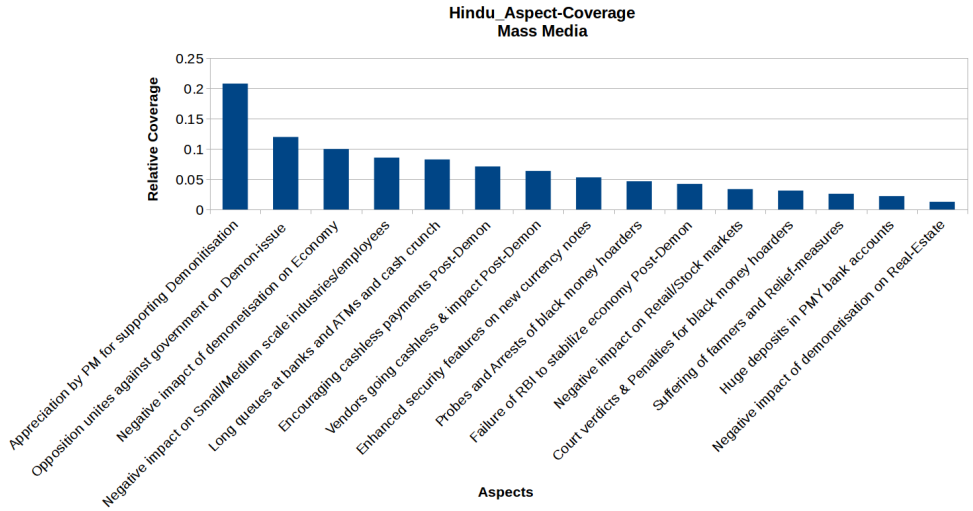


(a) TeleG

Figure A.6 [RQ1] Newspaper wise Relative Aspect coverage for Demonetisation in mass media.

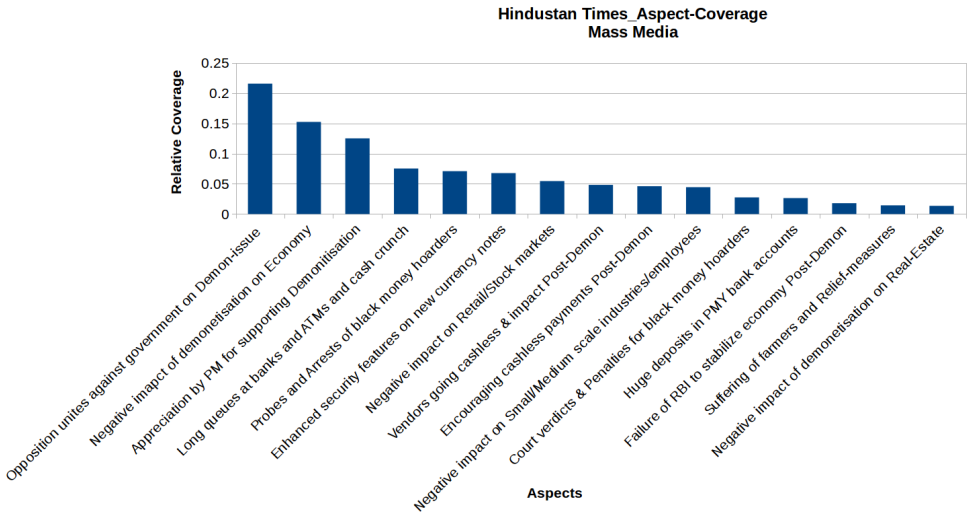


(a) TOI

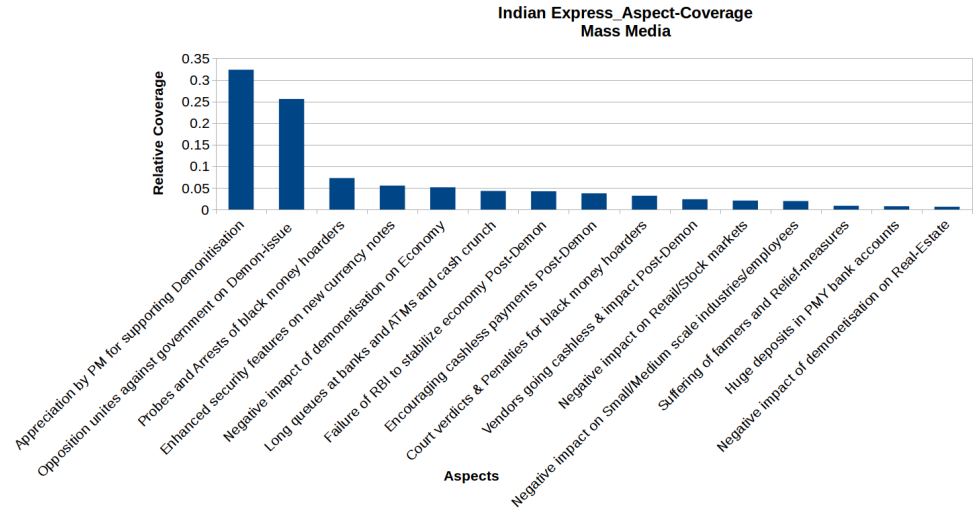


(b) The Hindu

Figure A.7 [RQ1] Newspaper wise Relative Aspect coverage for Demonetisation in mass media.

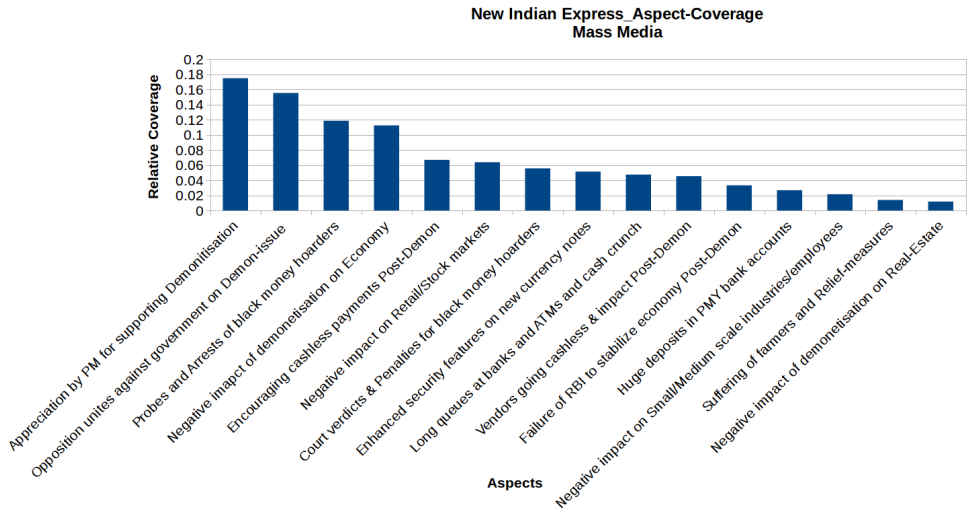


(a) HT

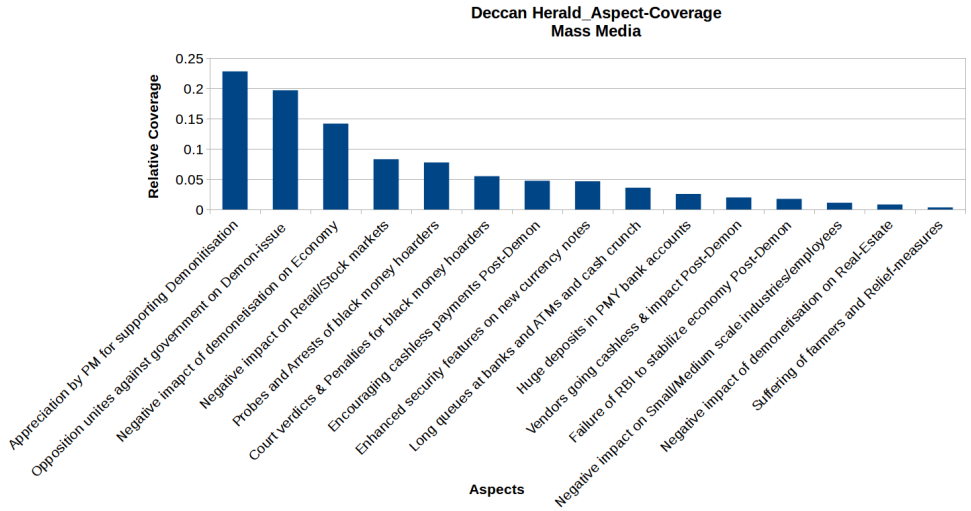


(b) IE

Figure A.8 [RQ1] Newspaper wise Relative Aspect coverage for Demonetisation in mass media.



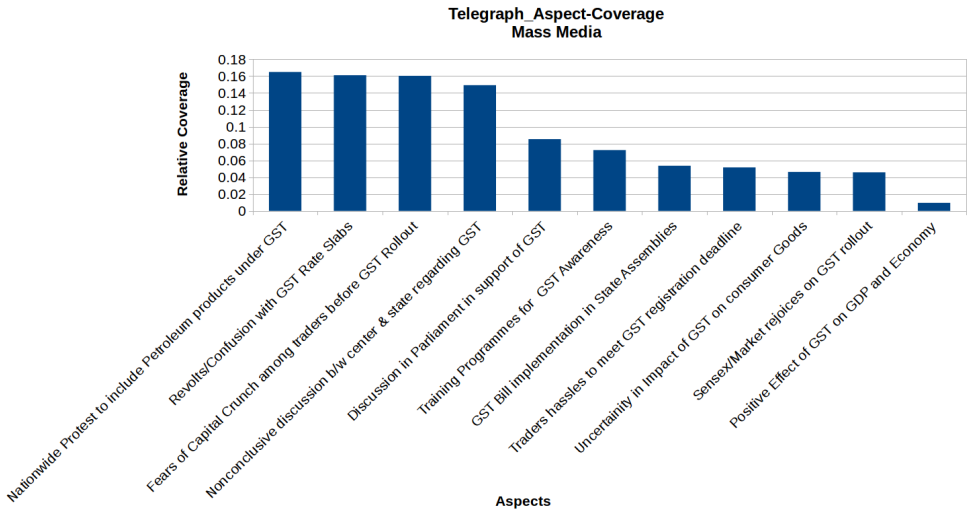
(a) NIE



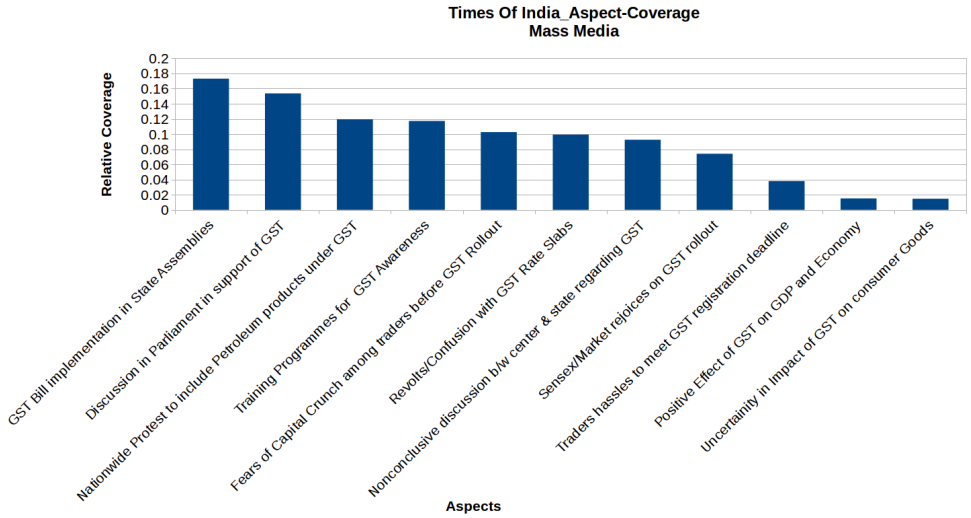
(b) DecH

A.6 Aspect Coverage for GST:

Figure A.9 [RQ1] Newspaper wise Relative Aspect coverage for GST in mass media.

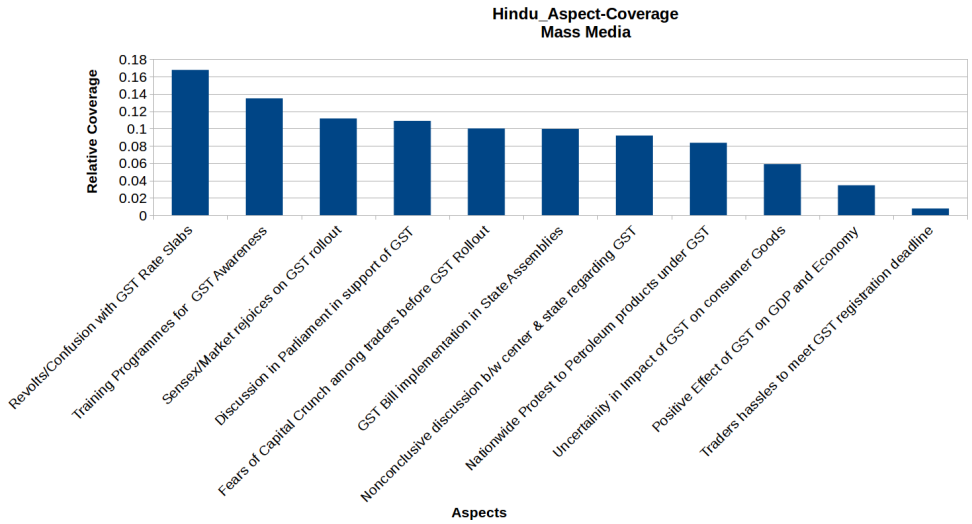


(a) TeleG

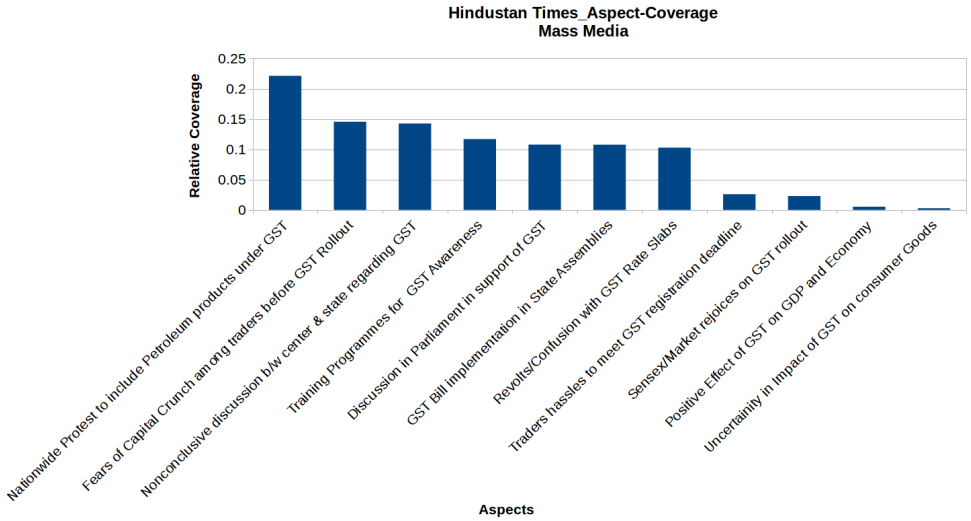


(b) TOI

Figure A.10 [RQ1] Newspaper wise Relative Aspect coverage for GST in mass media.

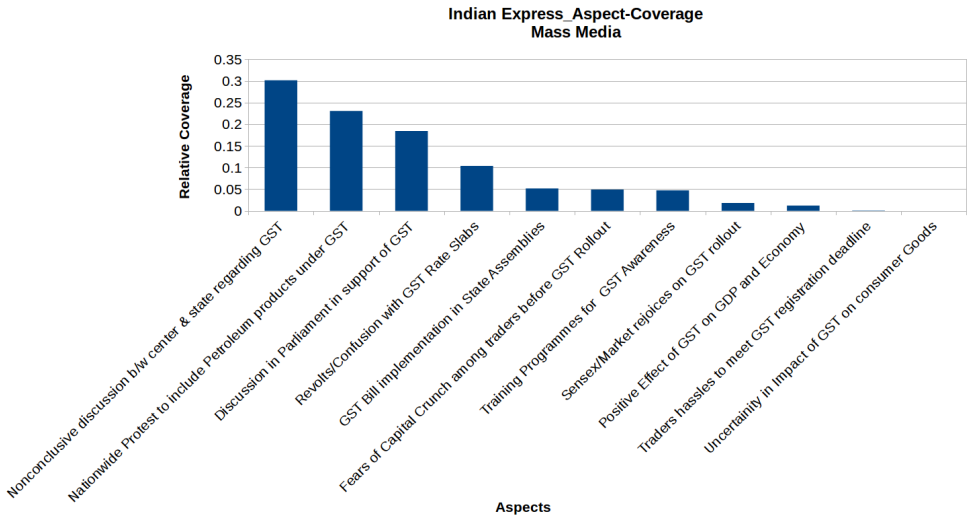


(a) The Hindu

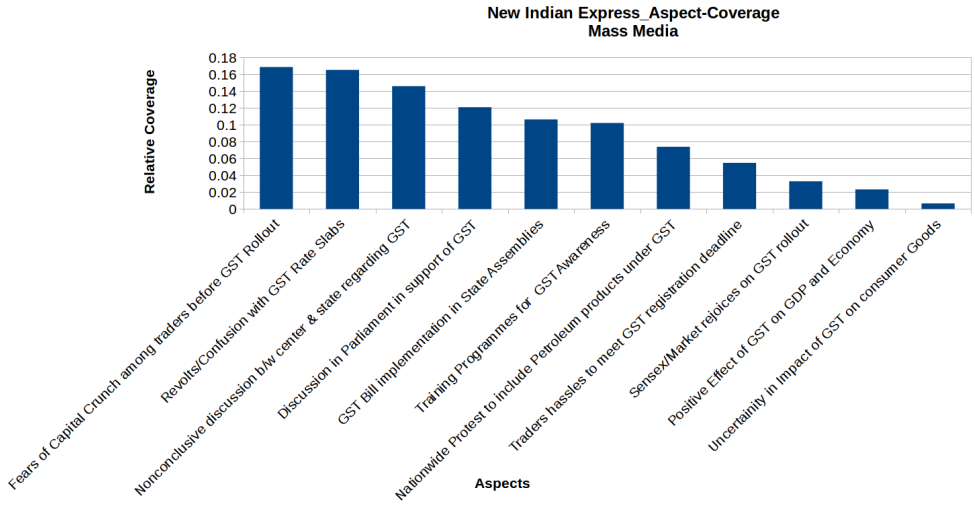


(b) HT

Figure A.11 [RQ1] Newspaper wise Relative Aspect coverage for GST in mass media.

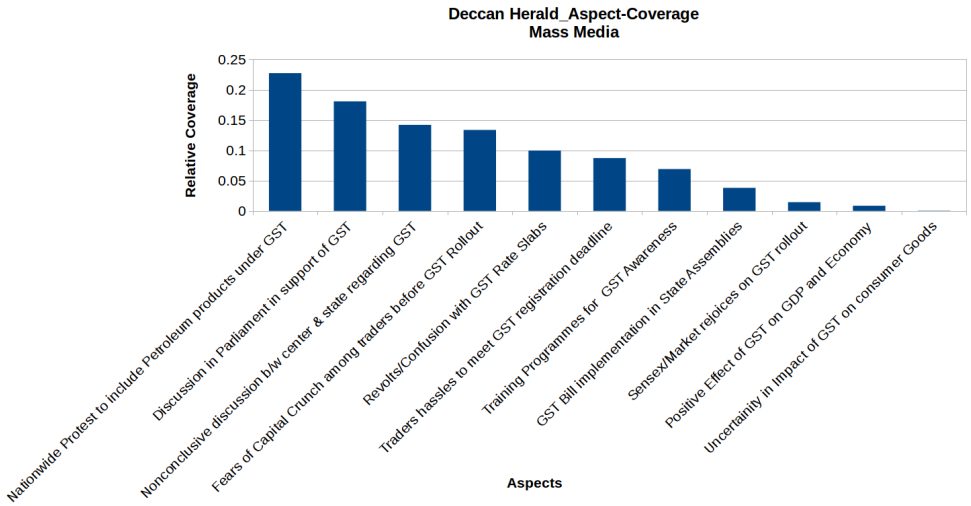


(a) IE



(b) NIE

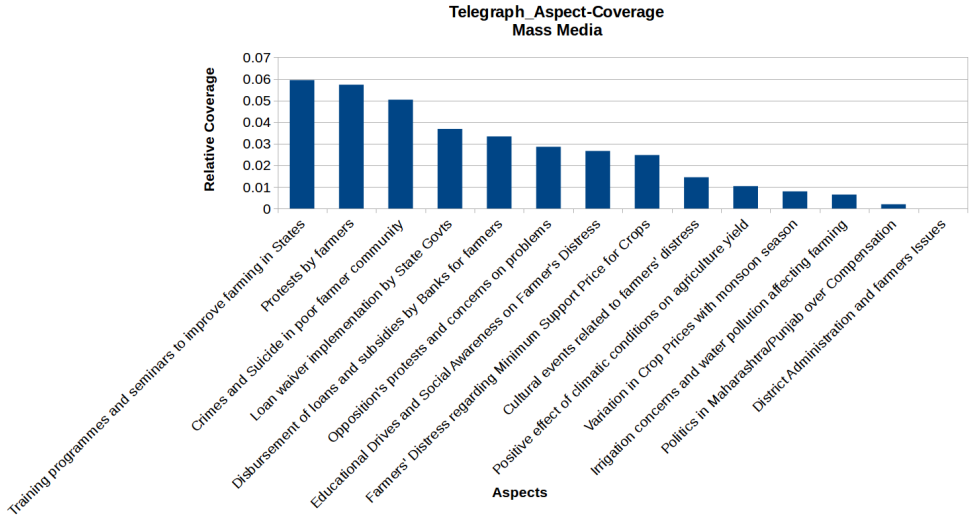
Figure A.12 [RQ1] Newspaper wise Relative Aspect coverage for GST in mass media.



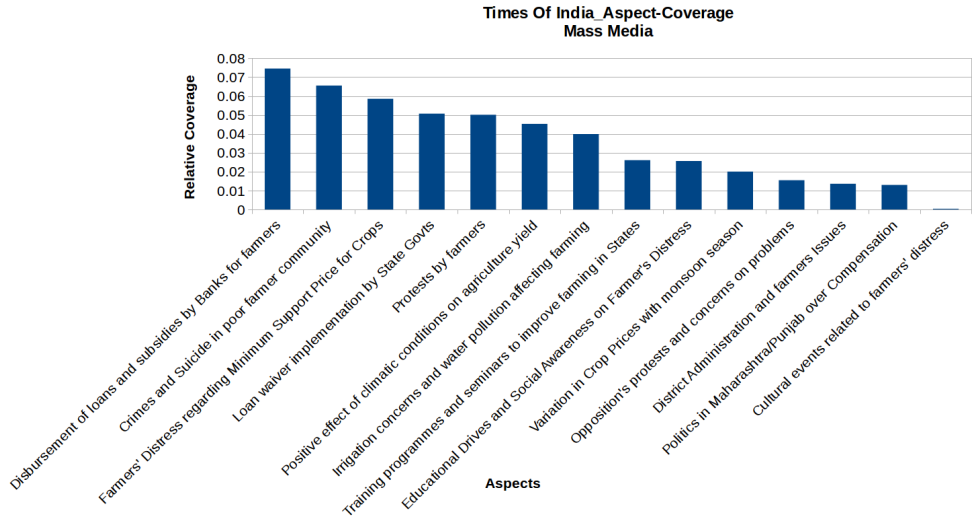
(a) DecH

A.7 Aspect Coverage for Farmer's Protest:

Figure A.13 [RQ1] Newspaper wise Relative Aspect coverage for Farmer's Protest in mass media.

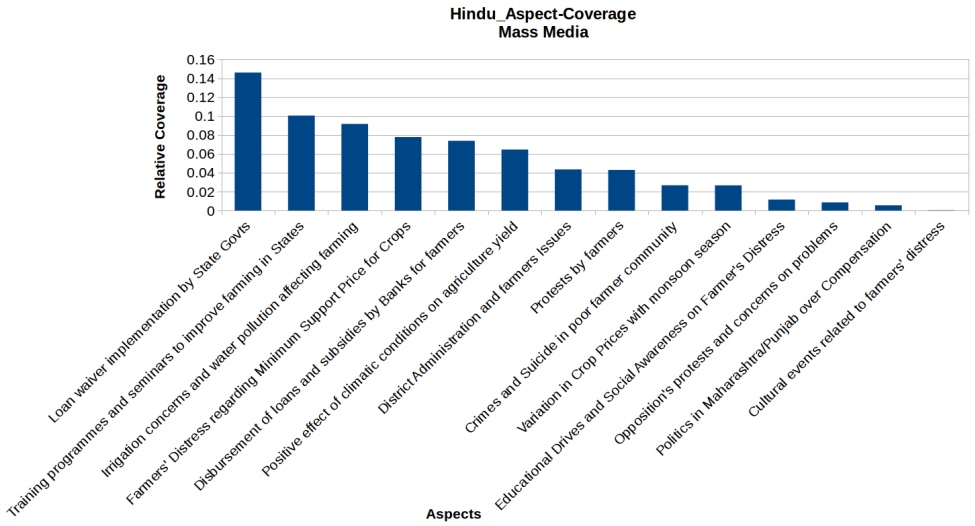


(a) TeleG

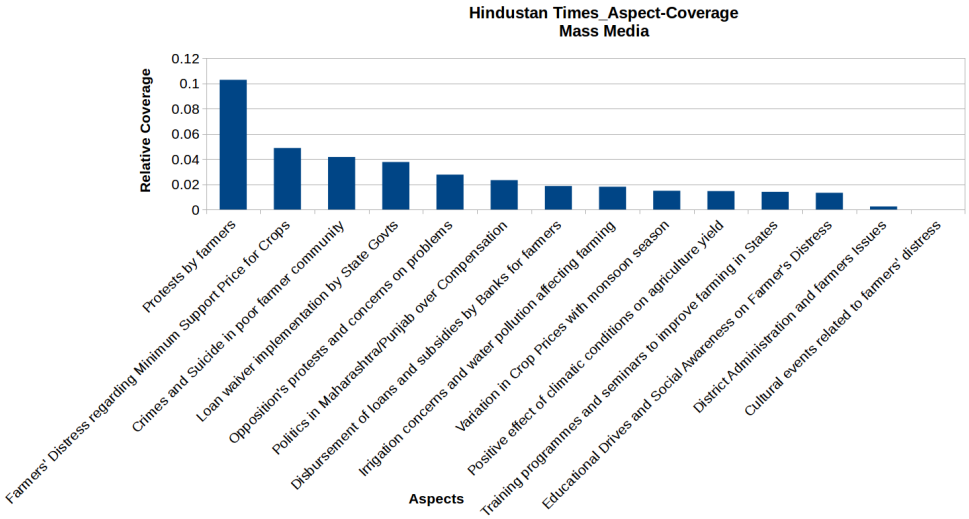


(b) TOI

Figure A.14 [RQ1] Newspaper wise Relative Aspect coverage for Farmer's Protest in mass media.

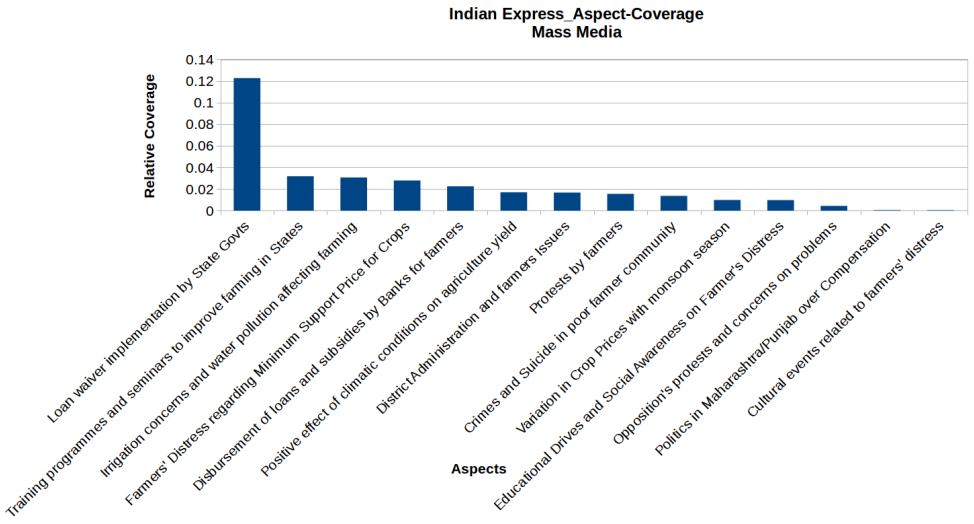


(a) The Hindu

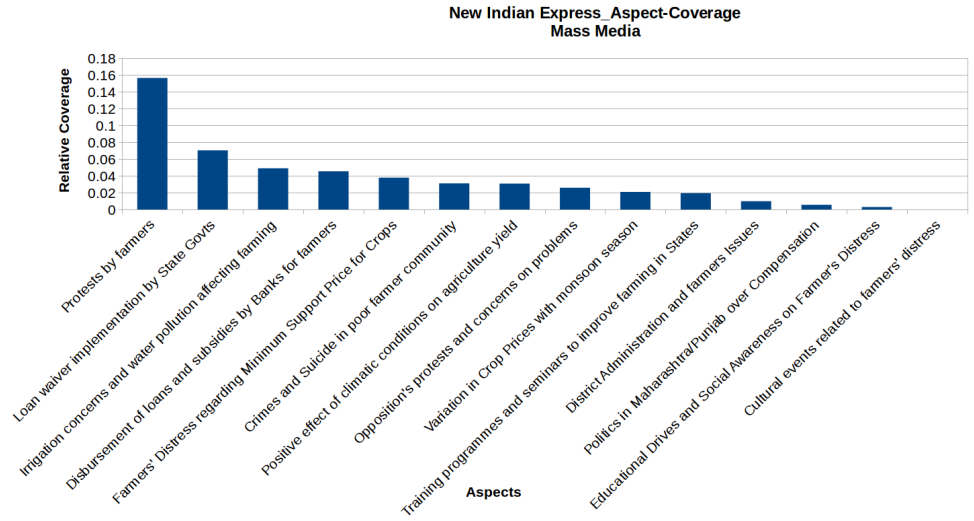


(b) HT

Figure A.15 [RQ1] Newspaper wise Relative Aspect coverage for Farmer's Protest in mass media.

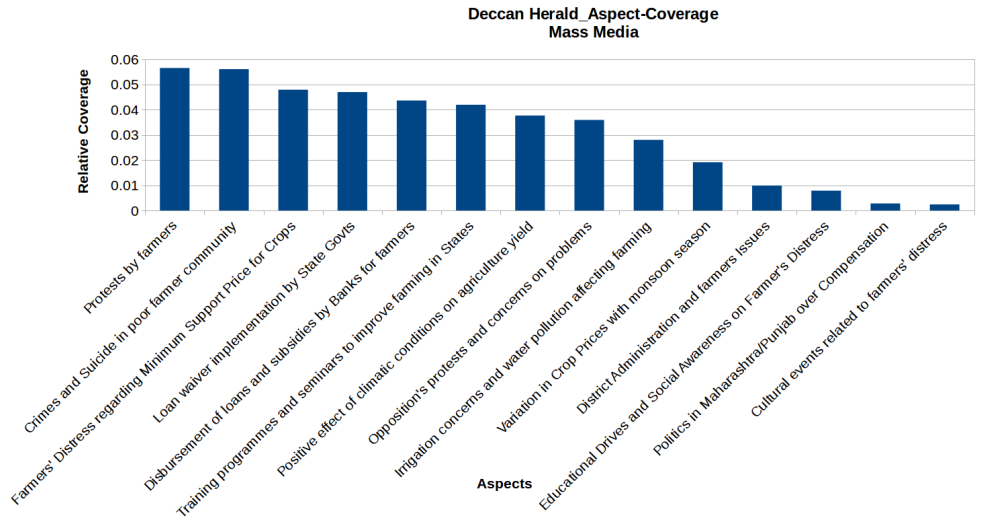


(a) IE



(b) NIE

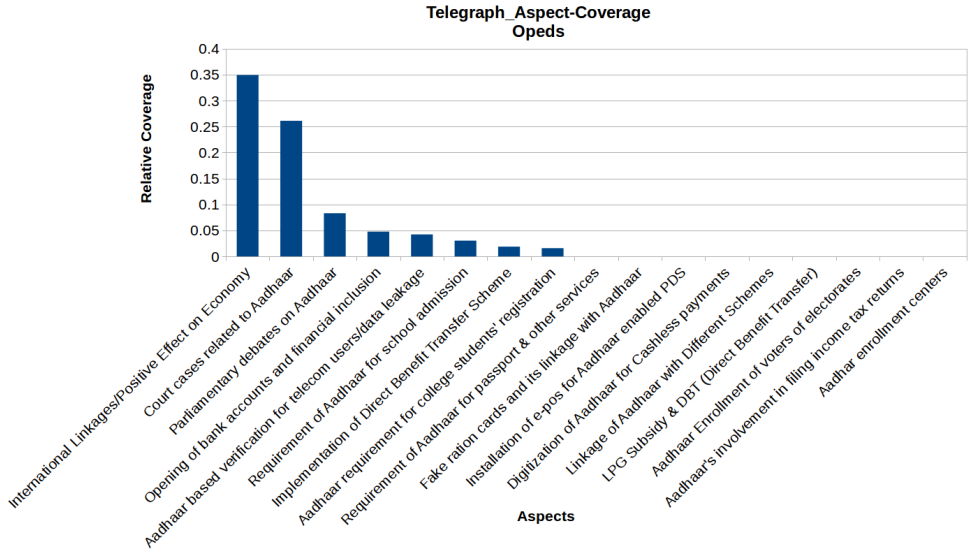
Figure A.16 [RQ1] Newspaper wise Relative Aspect coverage for Farmer's Protest in mass media.



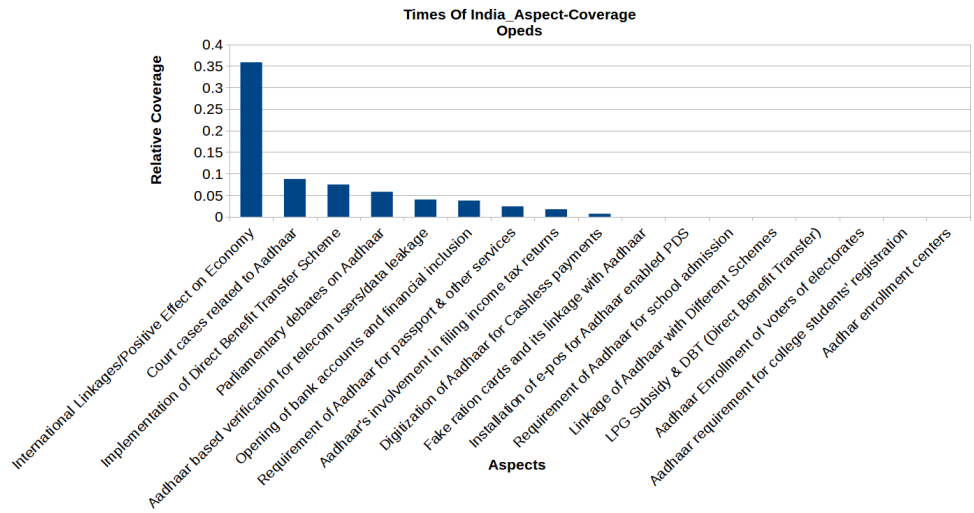
(a) DecH

A.8 Opeds Aspect Coverage for Aadhaar:

Figure A.17 [RQ7] Newspaper wise Relative Aspect coverage for Aadhaar in mass media for Opeds.

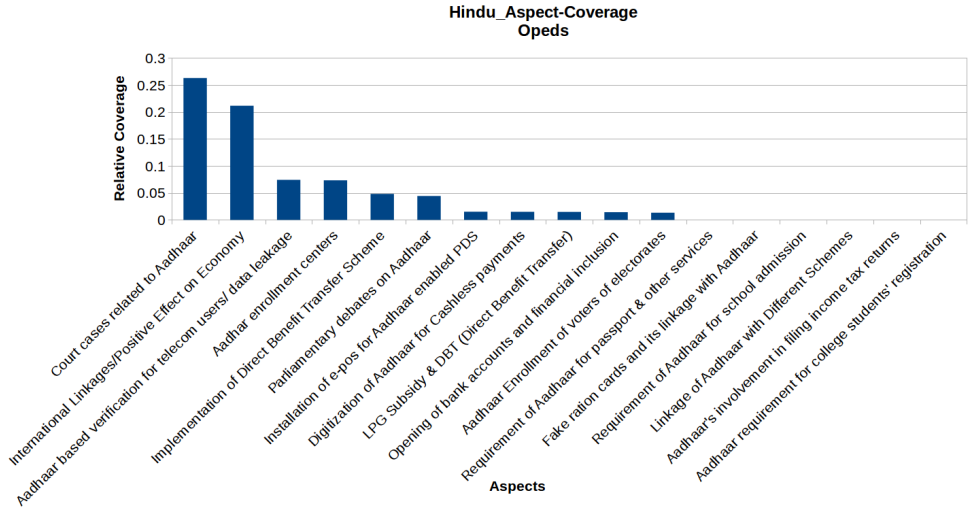


(a) TeleG

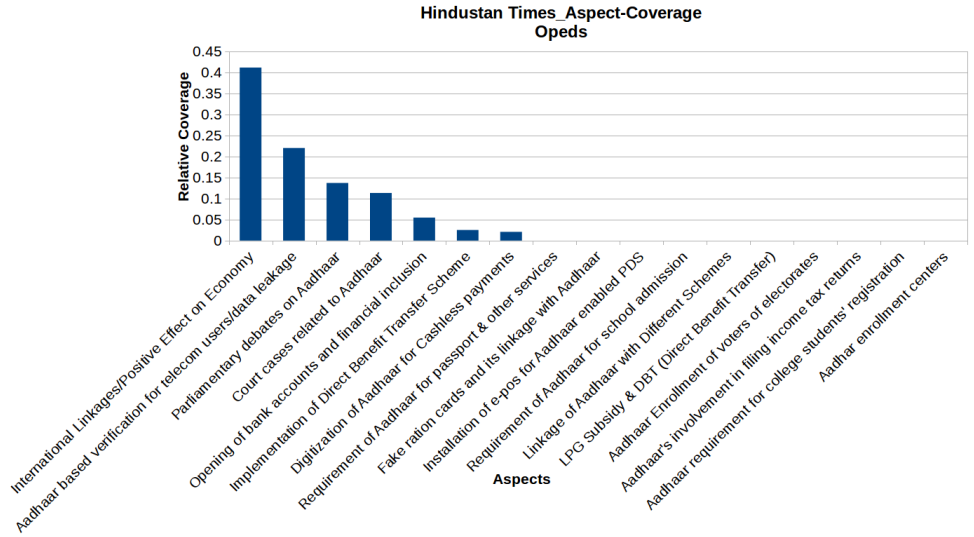


(b) TOI

Figure A.18 [RQ7] Newspaper wise Relative Aspect coverage for Aadhaar in mass media for Opeds.

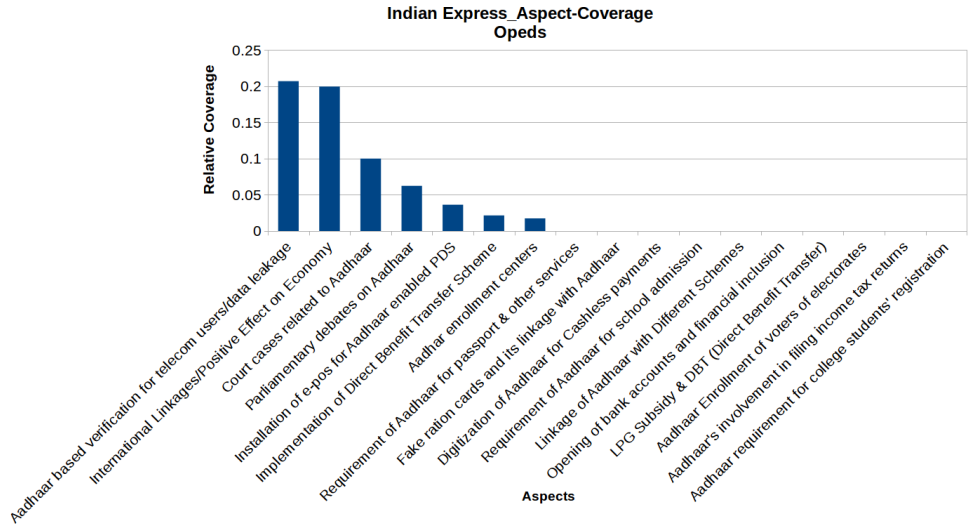


(a) The Hindu

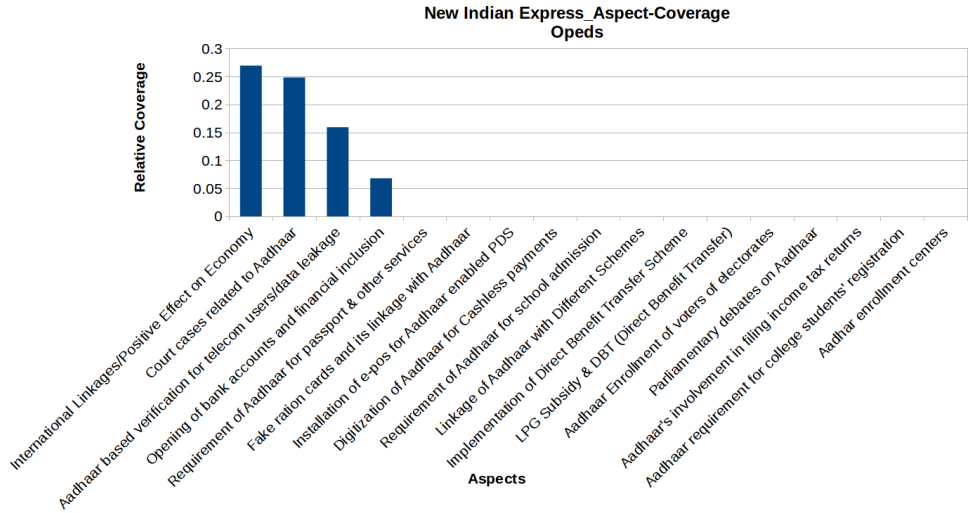


(b) HT

Figure A.19 [RQ7] Newspaper wise Relative Aspect coverage for Aadhaar in mass media for Opeds.

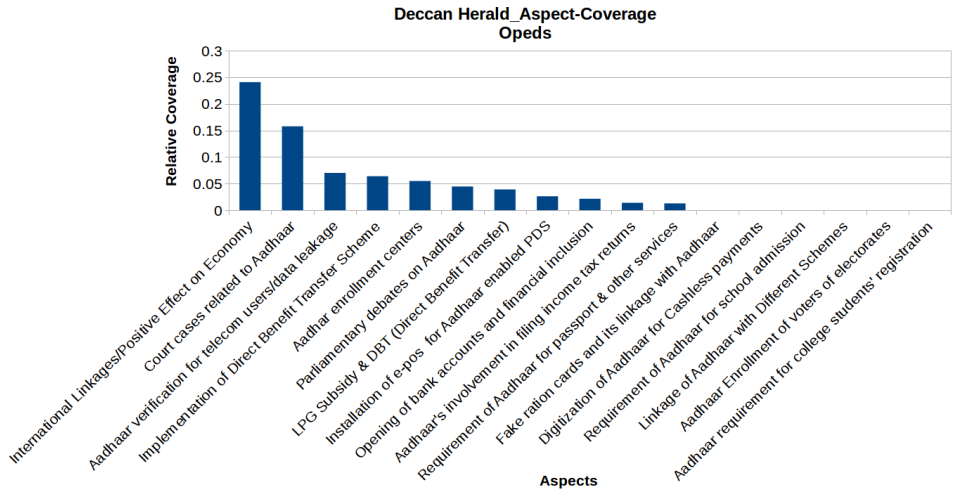


(a) IE



(b) NIE

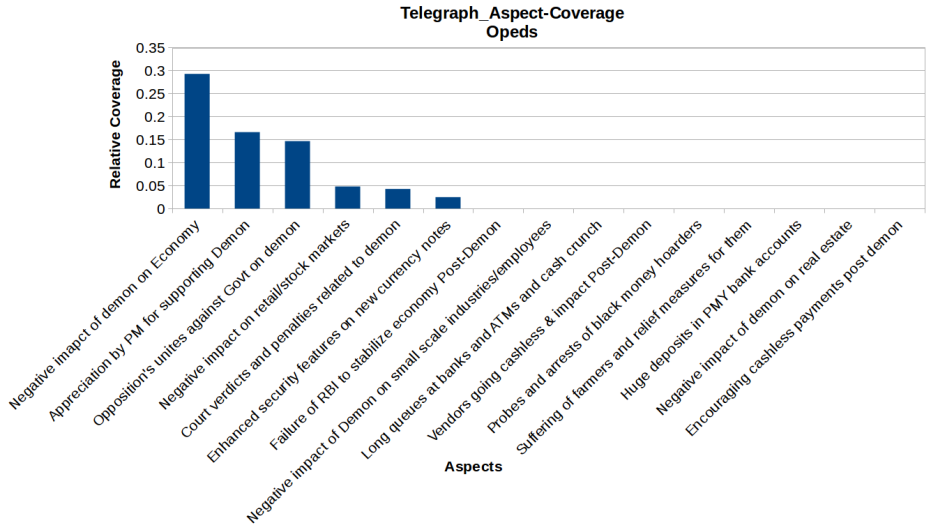
Figure A.20 [RQ7] Newspaper wise Relative Aspect coverage for Aadhaar in mass media for Opeds.



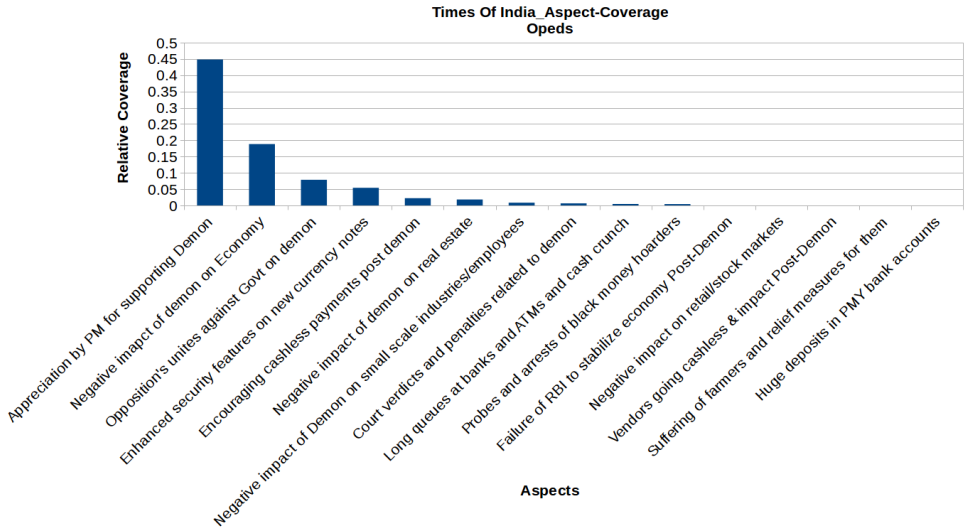
(a) DecH

A.9 Opeds Aspect Coverage for Demon:

Figure A.21 [RQ7] Newspaper wise Relative Aspect coverage for Demonetisation in mass media for Opeds.

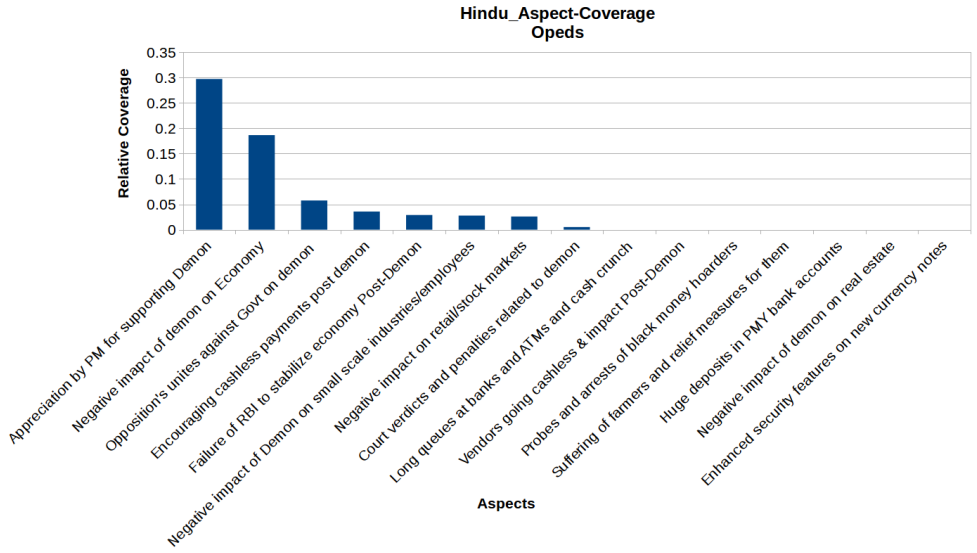


(a) TeleG

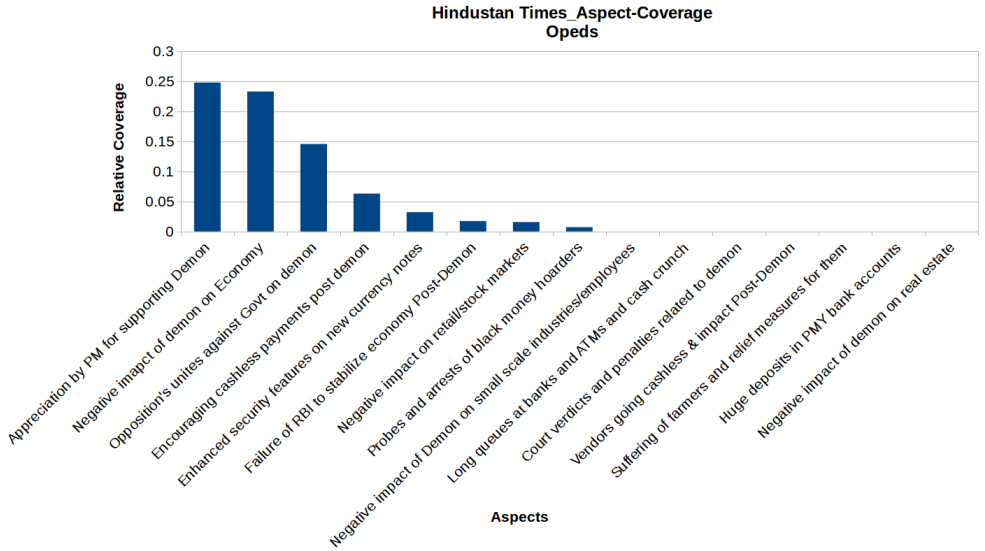


(b) TOI

Figure A.22 [RQ7] Newspaper wise Relative Aspect coverage for Demonetisation in mass media for Opeds.

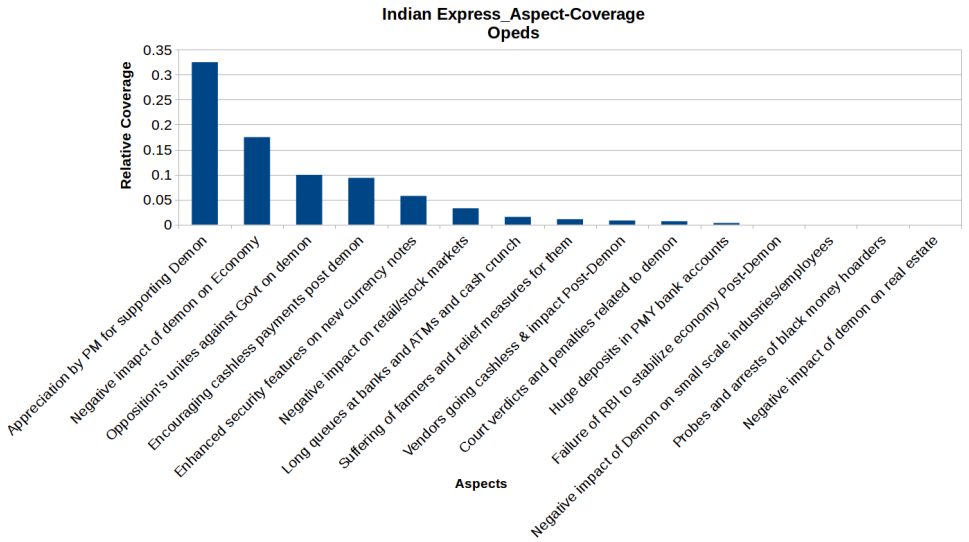


(a) The Hindu

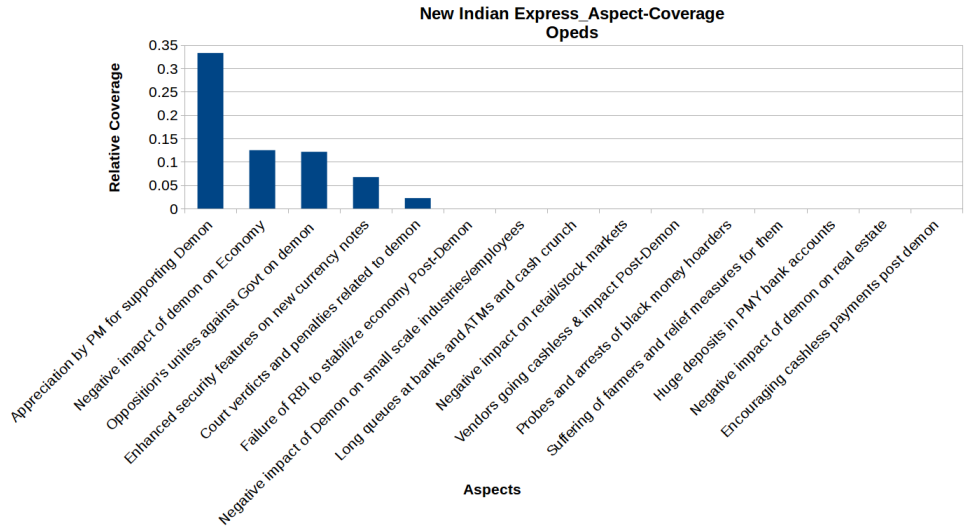


(b) HT

Figure A.23 [RQ7] Newspaper wise Relative Aspect coverage for Demonetisation in mass media for Opeds.

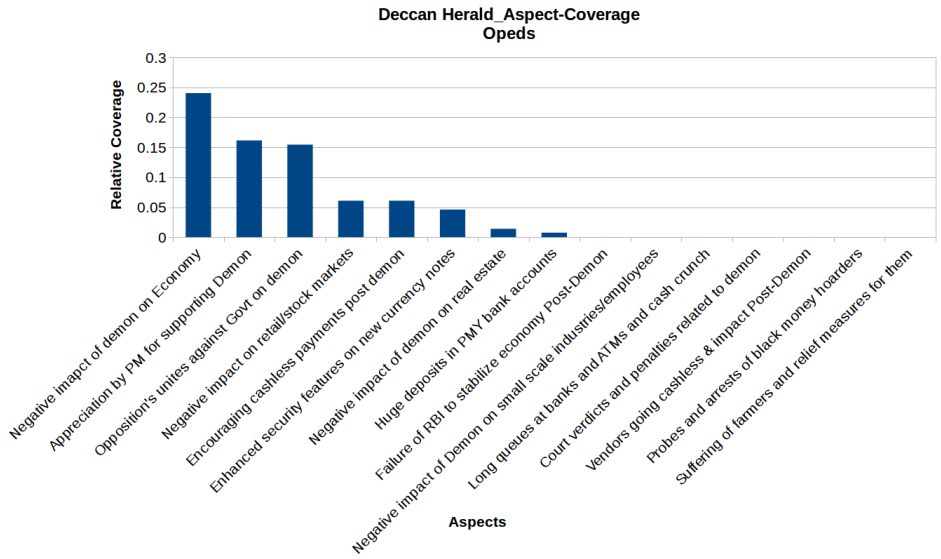


(a) IE



(b) NIE

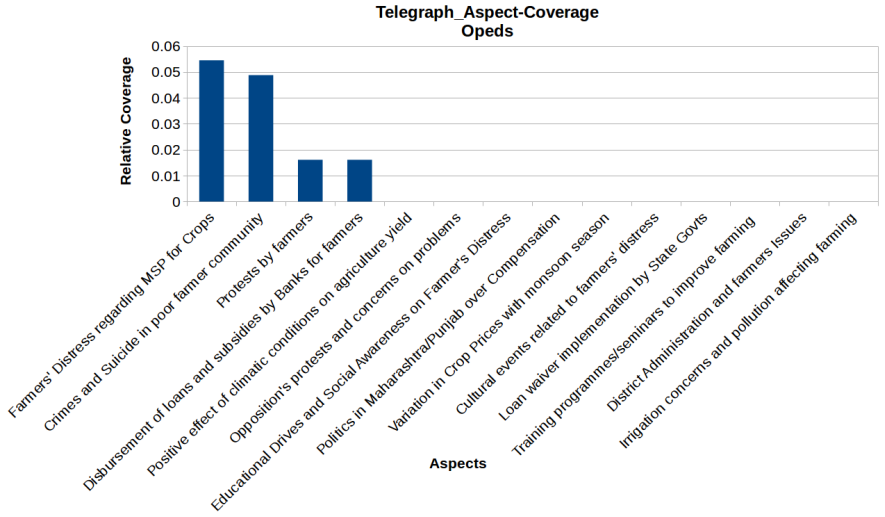
Figure A.24 [RQ7] Newspaper wise Relative Aspect coverage for Demonetisation in mass media for Opeds.



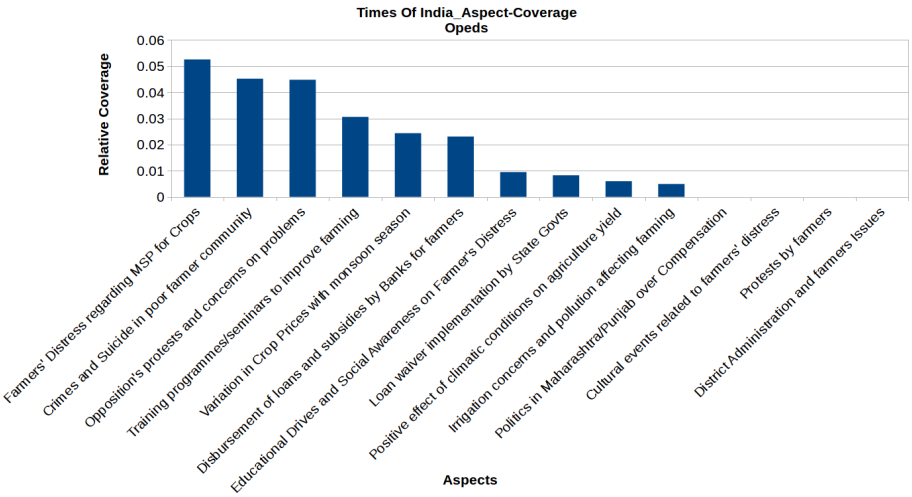
(a) DecH

A.10 Opeds Aspect Coverage for FP:

Figure A.25 [RQ7] Newspaper wise Relative Aspect coverage for Farmer's Protest in mass media for Opeds.

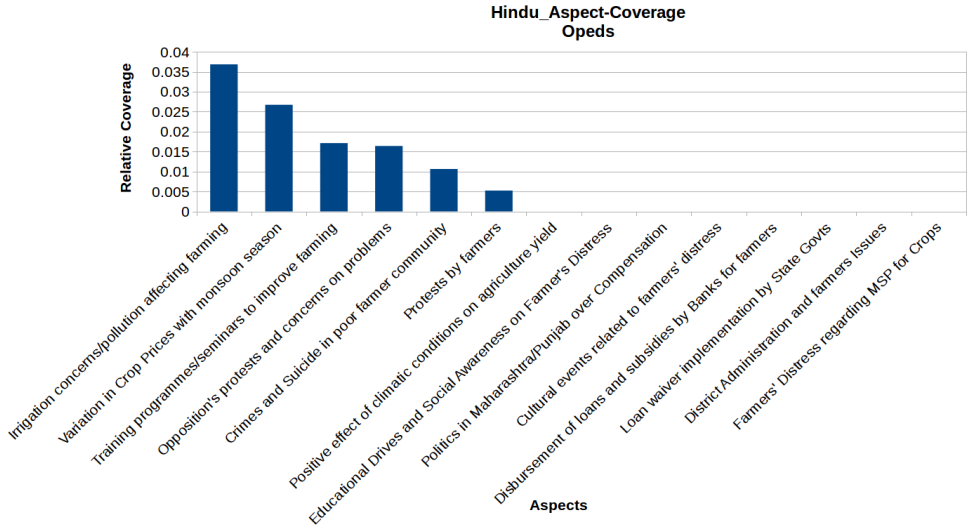


(a) TeleG

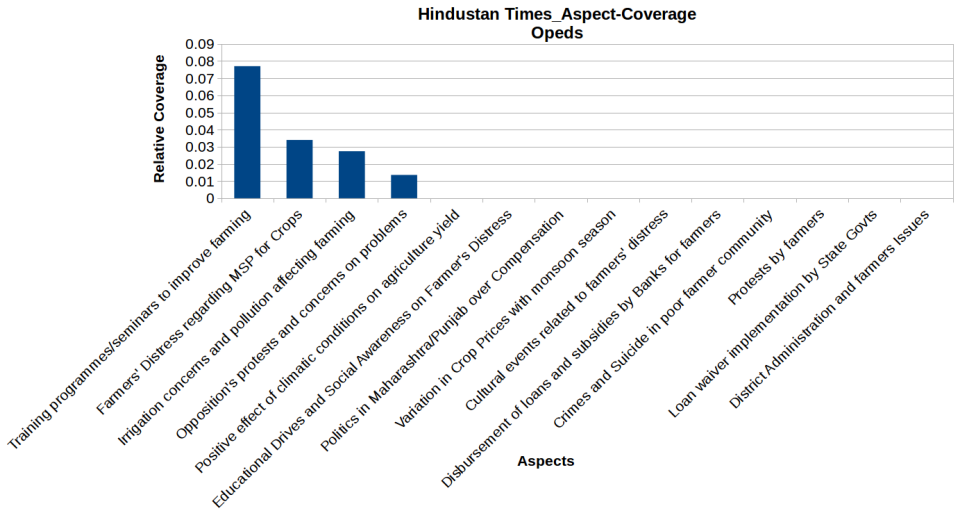


(b) TOI

Figure A.26 [RQ7] Newspaper wise Relative Aspect coverage for Farmer's Protest in mass media for Opeds.

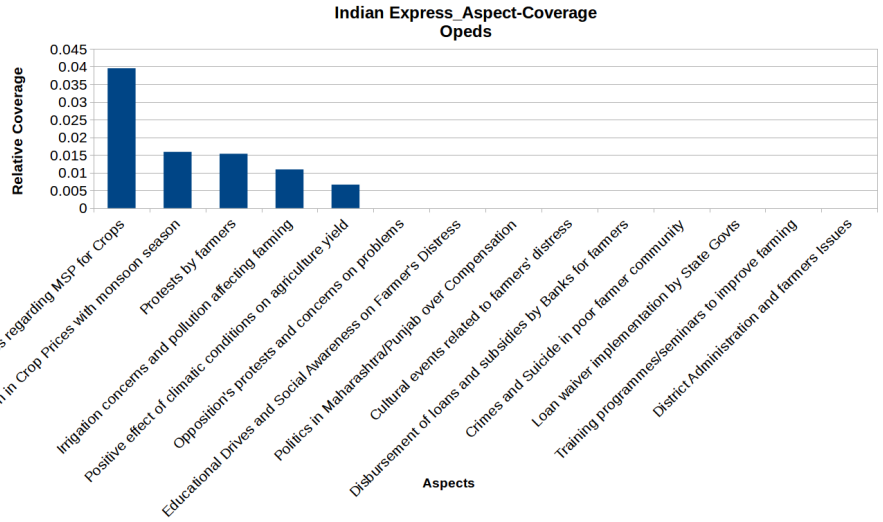


(a) The Hindu

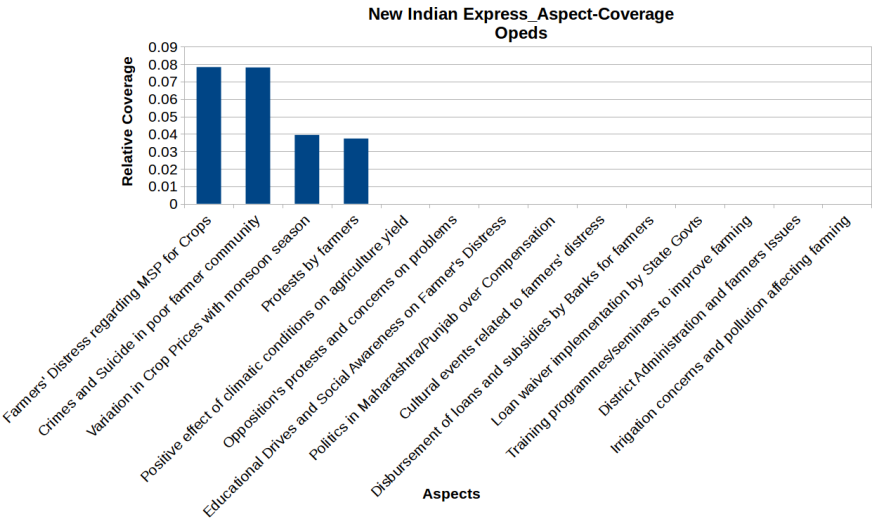


(b) HT

Figure A.27 [RQ7] Newspaper wise Relative Aspect coverage for Farmer's Protest in mass media for Opeds.

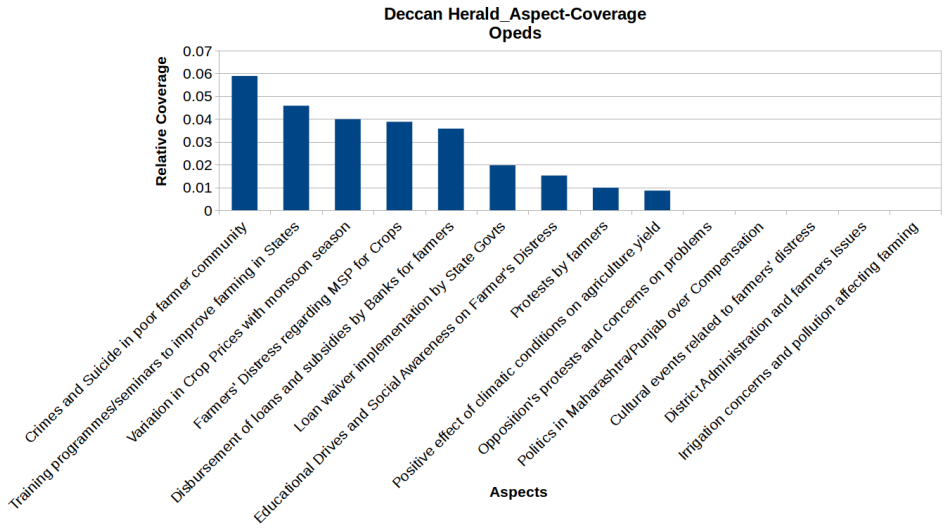


(a) IE



(b) NIE

Figure A.28 [RQ7] Newspaper wise Relative Aspect coverage for Farmer's Protest in mass media for Opeds.



(a) DecH

References

[1] Stephen D Reese, Tim P Vos, and Pamela J Shoemaker. 2009. Journalists as gatekeepers. In The handbook of journalism studies. Routledge, 93107.

[2] Chun-Fang Chiang and Brian Knight. 2011. Media bias and influence: Evidence from newspaper endorsements. The Review of Economic Studies 78, 3 (2011),795820.

[3] Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? Evidence from US daily newspapers. Econometrica 78, 1 (2010), 3571.

[4] Filipe N Ribeiro, Lucas Henriqueo, Fabricio Benevenutoo, Abhijnan Chakraborty,Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna P Gummadi. 2018. Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale.(2018).

[5] Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. Public Opinion Quarterly 80,S1 (2016), 250271.

[6] Sendhil Mullainathan and Andrei Shleifer. 2005. The market for news. American Economic Review 95, 4 (2005), 10311053.

[7] Andrea Prat and David Strmberg. 2013. The political economy of mass media.Advances in economics and econometrics 2 (2013), 135.

[8] David Strmberg. 2004. Mass media competition, political competition, and public policy. The Review of Economic Studies 71, 1 (2004), 265284.

[9] Larry M Bartels. 1993. Messages received: The political impact of media exposure.American Political Science Review 87, 2 (1993), 267285.

[10] Paul F Lazarsfeld, Bernard Berelson, and Hazel Gaudet. 1944. The peoples choice: How the voter makes up his mind in a presidential election. New York:Duell, Sloan and Pearce (1944).

[11] Dietram A Scheufele and David Tewksbury. 2006. Framing, agenda setting, and priming: The evolution of three media effects models.Journal of communication 57, 1 (2006), 920

[12] Pamela E Oliver and Daniel J Myers. 1999. How events enter the public sphere:Conflict, location, and sponsorship in local newspaper coverage of public events.American journal of sociology 105, 1 (1999), 3887.

[13] Edward S Herman. 1988. 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.).

[14] Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? Evidence from US daily newspapers.Econometrica 78, 1 (2010), 3571.

[15] Sean Gerrish and David M Blei. 2011. Predicting legislative roll calls from text.In Proceedings of the 28th international conference on machine learning (icml-11).489496.

[16] Yanchuan Sim, Brice DL Acree, Justin H Gross, and Noah A Smith. 2013. Measuring ideological proportions in political speeches. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.91101.

[17] Viet-An Nguyen, Jordan L Ying, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In Advances in neural information processing systems. 11061114.

[18] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1. 11131122.

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

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

[21] Lada A. Adamic and Natalie Glance. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In Proceedings of the 3rd International Workshop on Link Discovery (LinkKDD 05). ACM, New York, NY, USA, 3643.https://doi.org/10.1145/1134271.1134277

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

[23] Mark S Granovetter. 1977. The strength of weak ties. In Social networks. Elsevier,347367.

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

[25] Aaditeshwar Seth and Jie Zhang. 2008. A Social Network Based Approach to Personalized Recommendation of Participatory Media Content.. In ICWSM.

[26] Jisun An, Meeyoung Cha, P Krishna Gummadi, and Jon Crowcroft. 2011. Media Landscape in Twitter: A World of New Conventions and Political Diversity.. In ICWSM.

[27] Pablo Barber and Thomas Zeitzoff. 2017. The new public address system: why do world leaders adopt social media?International Studies Quarterly 62, 1 (2017),121130.

[28] Joyojeet Pal and Lia Bozarth. 2018. Is Tweeting in Indian Languages Helping Politicians Widen Their Reach? https://www.epw.in/engage/article/tweeting-indian-languages-helping-politicians-widen-reach.

[29] JOYOJEET PAL, PRIYANK CHANDRA, PADMA CHIRUMAMILLA, VAISHNAV KAMESWARAN, ANDRE GONAWELA, UDIT THAWANI, and PRITIKA DASGUPTA. 2017. Innuendo as Outreach:@ narendramodi and the Use of Political Irony on Twitter.International Journal of Communication (19328036) 11 (2017).

[30] Joyojeet Pal, Udit Thawani, Elmer Van Der Vlugt, Wim Out, Priyank Chandra,et al. 2018. Speaking their Mind: Populist Style and Antagonistic Messaging in the Tweets of Donald Trump, Narendra Modi, Nigel Farage, and Geert Wilders.Computer Supported Cooperative Work (CSCW) 27, 3-6 (2018), 293326.

[31] Joyojeet Pal. 2019. Legitimacy, Support and Endorsement: Narendra Modis Social Media Engagement with Celebrities. https://www.epw.in/engage/article/legitimacy-support-and-endorsement.

[32] Joyojeet Pal, Priyank Chandra, and VG Vinod Vydiswaran. 2016. Twitter and the rebranding of Narendra Modi.Economic Political Weekly 51, 8 (2016), 5260.

[33] Karen Bird. 2005. Gendering parliamentary questions.The British Journal of Politics and International Relations 7, 3 (2005), 353370.

[34] Srikrishna Ayyangar and Suraj Jacob. 2015. Question Hour Activity and Party Behaviour in India. The Journal of Legislative Studies 21, 2 (2015), 232249.

[35] Kelly Blidook and Matthew Kerby. 2011. Constituency influence on constituency members: The adaptability of roles to electoral realities in the Canadian case.The Journal of Legislative Studies 17, 3 (2011), 327339.

[36] Stefanie Bailer. 2011. Peoples voice or information pool? The role of, and reasons for, parliamentary questions in the Swiss parliament.The Journal of Legislative Studies 17, 3 (2011), 302314.

[37] Thomas Saalfeld. 2011. Parliamentary questions as instruments of substantive representation: Visible minorities in the UK House of Commons, 200510.The Journal of Legislative Studies 17, 3 (2011), 271289.

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

[39] Jeremy Holland. 2007. Tools for institutional, political, and social analysis of policy reform: a sourcebook for development practitioners. The World Bank.

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

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

[42] Anirban Sen, A Agarwal, Aditya Guru, A Choudhuri, G Singh, Imran Mohammed,J Goyal, K Mittal,Manpreet Singh, Mridul Goel, et al. 2018. Leveraging Web Data to Monitor Changes in Corporate-Government Interlocks in India. In Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies.ACM, 15.

[43] Anirban Sen and ACT4D Team. 2018. Supplementary Information: An Attempt at Using Mass Media Data to Analyze the Political Economy Around Some Key ICTD Policies in India. http://bit.ly/2OqFgme.

[44] CJ Hutto Eric Gilbert. 2014. 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.

[45] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas.2010. Sentiment in Short Strength Detection Informal Text. J. Am. Soc. Inf. Sci.Technol. 61, 12 (Dec. 2010), 25442558. https://doi.org/10.1002/asi.v61:12

[46] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard,and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. 5560.