An Empirical Analysis of Media Bias in India on the Coverage of Economic Policy Issues

Content authored by news-sources is a significant source of web content. This content however reflects the biases in mass media, which can be amplified further through various algorithms for search and recommendation that operate on web data. Digital governance tools that reveal the biases in content published by news-sources can provide important hints to algorithm developers, and also serve as a self-regulatory mechanism for the media to reflect upon its biases. We examine the coverage given on the Indian mass media and social media to four significant economic policy issues, and build methods to quantify biases of different news-sources. Other than being one of the first large scale studies in the Indian context, our work is significant in creating a standardized methodology to assess the ideological slant of a news-source, and its alignment with the social media discourse of the follower community of the news-source. We have created a web-based platform to disclose these biases of different news-sources, and trust that it can be a step towards digital governance of bias in web content.

CCS Concepts: • **Information systems** → **Information systems applications**; • **Social and professional topics** → *Computing / technology policy*; • **Applied computing** → Computers in other domains;

Additional Key Words and Phrases: Bias on the web, digital governance, digital divide, mass media bias, web data, social media analysis, media bias monitor, sentiment analysis, social media community

1 INTRODUCTION

Analyzing the existence of biases in web data has been an active area of research [9, 10, 20, 55]. In today's age of personalization of web content, recommendation systems and search engines can further reinforce these biases through their algorithms [23, 54]. Such biases in web content inherently arise due to demographic inequalities, differences in political ideologies, and similar factors that give differential opportunities to different users to produce content. In this paper, we examine one type of web content, that produced by news sources which regularly create new content and publish it on the web. This content by news sources could be biased due to various factors that influence media bias, and we build a set of tools to quantify this bias. Through an analysis of seven prominent national English news sources in India on their coverage of four recent economic policies, we show that indeed the Indian mass media has certain biases, that the biases seem to be aligned with biases in social media content, and that this can continue to perpetuate biases and inequalities in web content. Based on the tools we have developed, we have built an online platform to serve as a self-regulatory mechanism for the mass media in India, towards potentially addressing this bias in the production of new content.

News media is known to inform public opinion on different policies towards shaping how people think about these policies, and even what specific aspects of the policies do people think about [28]. Investigating the bias in how policies are represented in the media is therefore an important area of study. Biases in media can take different forms such as coverage bias on how much attention is

XXXX-XXXX/2019/8-ART1
https://doi.org/10.1145/nnnnnnn.nnnnnnn

given to a policy or to different aspects about a policy, selection bias on the amount of coverage given to different people or political parties, and sentiment bias on how positively or negatively different aspects and entities are represented in media [46].

These biases influence the priorities that the readers place on various policies, or aspects related to the policies. The biases in mass media, along with the impact they have on the prioritization of issues by the readers, is called *agenda setting*. To study this effect, we analyze in detail these different forms of biases on how four recent important issues are represented in seven leading English national news-sources in India, and compare the mass media coverage with social media content on these policy issues to determine the concordance between these two spheres of expression of public opinion. We also study the effect of *framing* in Indian mass media, by seeing how the newspapers talk about the policy from the perspective of five different constituencies of the *poor, middle class, informal sector, corporate, and government.* These five constituencies are the audience towards whom the mass media presents information about the policies considered. It must be noted that these *constituencies* are not the same as political *ideologies.* A constituency focus simply tells us if a news-source presents its content using a particular frame, such as talking positively or negatively about the effect of the policies on the poor, or on the middle class, etc.

Biases in mass media could arise due to several reasons, like definite slants of different news-sources based on constituencies (for example pro-poor, pro-consumers, pro-corporate, etc.), political affiliations of news-sources (towards different political parties), and may even be driven by commercial factors to fine-tune the coverage and bring it in sync with the preferences of their readers [26, 35]. These biases may even be dynamic and affected by changing factors like ownership networks of media companies [14, 52], growing concentration in the media industry [7, 15], and polarization of audiences [39]. Both sociological models [25] and economic models [34] have been proposed to explain these observations, but have not been strongly validated through large datasets. In this paper, other than giving a view of bias in mass media, our key contribution is also a computational framework to analyze the coverage of any news event in the mass media and social media, and can be adapted to similarly analyze other sources of web content.

We consider four prominent economic policy issues, namely *Demonetization* [17], *Aadhaar* [4], *GST* [57], and *Farmers' Protest* [56], each of which is a very recent and actively debated policy issue. We describe them briefly here: **(a) Demonetization**: 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, etc. **(b) Aadhaar**: An initiative by the government to give every Indian resident a biometric-based unique identification number. 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. **(c) 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 significant involvement of different politicians and political parties. **(d) Goods and Services Tax**: 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 debates though on its complexity and problems in implementation which have impacted the overall growth of the economy.

For each of these policy events, we extract a set of *aspects* (commonly discussed topics) from the articles published by news-sources using an unsupervised topic-modeling method called Latent

Dirichlet Allocation (LDA)[11]. Our analysis of these four policy issues and their aspects helps us build strong evidence about whether or not the mass media is biased. We analyze the following questions:

- **Ideological slant of news-sources**: (a) Are news sources biased on the amount of coverage they give to different aspects about the policy issues? (b) Do news-sources have a bias towards or against constituencies like the poor, middle class, government, informal sector, and corporate?
- **Alignment of news-sources with their audience**: Are some news-sources more closely aligned with their viewers (on social media) than others?

Overall, we find that: (a) Ideological bias does exist w.r.t. the coverage provided to different aspects, (b) news-sources differ in terms of their alignment towards different constituencies, and the methods they use to convey their slants, with most news-sources being more aligned to middle class and government constituencies, providing comparatively lesser coverage to issues related to the poor, and (c) the news-sources seem to be strongly aligned with their follower community on Twitter. These findings together provide strong hints towards the agenda setting and framing effects exercised by the Indian mass media that might lead towards influencing public opinion on key policies.

Our findings are indicative of biases in web based media content and social network content, in terms of the coverage and slant given to different aspects around the policies. Given these biases, we conclude that (a) the digital divide persists in the coverage of issues in both mass media and social media, where the issues of the poor are much less covered, while significant coverage is provided to political issues and the issues related to the middle class, and (b) echo chambers seem to exist in social media, and the social media readership is strongly aligned to their favorite media houses in terms of the issues that they follow. Our key contribution is the generic framework, which can be applied on more topics to study the bias of data in web based sources and social media, and serve to provide indicators to algorithm developers of search engines and content recommendation platforms, about the biases embedded in web based content. We are currently developing a recommendation system, which uses the methods described in this paper to audit news feeds, and generates a news feed which is fair and unbiased, with a view to counter the bias that already exists in the web content in media.

2 RELATED WORK

We divide studies related to media bias into two parts: (a) bias in the web and social media, and (b) bias in the mass media.

Bias in the web and social media: 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 [8]. Garimella et al. in their paper [20] study the polarization of users on Twitter in terms of the content they post on controversial debates, and find that this polarization increases with the increase in interest about the event. Eli Pariser [40] discusses the issue of personalization on the Internet through Google's segregation of its user base into different filter bubbles. News-sources and influential journalists can impact audience attention highly through social media as well, thereby propagating their inherent bias [38]. Bias is also reflected in the web browsing history of users as studied in the US scenario by Flaxman et al. [18] - the authors find that most people visit a handful of ideologically similar news outlets, leading to less diversity. Similar biases have been noticed in the linking pattern of social media such as blogs [3] which leads to echo chambers. Such a view is corroborated by [41] 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 [24] and also noticed in [41, 48]. 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 [6].

Thus, there have been several studies that talk about algorithmic bias. In this work, however, we show that a significant bias exists in the content itself, both in terms of what issues (aspects) are covered, and with what sentiment. Since a significant amount of content on the web is produced by mass media, in this work we look at what kind of biases exist in the mass media web data. We therefore draw attention to the fact that not only is it required to deal with the algorithmic bias that exists with the search engines and social networks, but that self-regulatory mechanisms are needed as well for media institutions, so that there is less bias and adequate diversity in the content that they produce.

Bias in mass media: Journalists and news-sources shape public opinion by intentionally or inadvertently creating bias in their selection, writing, and distribution of news content, and for this reason they have often been called *gatekeepers* [42]. Scheufele et al. discuss the concepts of agenda setting, framing, and priming in mass media in their work [47]. These three effects together play a significant role in influencing public opinion on socio-political issues. Chiang et al. bring [13] out evidence of endorsements provided to political candidates by mass media in the USA. [21] 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. [44] 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.

There have been several studies on media bias assisted through computer-science techniques like [12, 27, 37]. Budak et al. [12] 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 quantifying the alignment of news-sources towards specific constituencies, and their agreement with their audience on social media. To the best of our knowledge, this is the first attempt to build a standard framework to specifically quantify news-sources on their slants based on constituencies and audience alignment that can be applied to any event.

Several media and web monitoring tools do currently exist. The Media Cloud platform [19], developed by Berkman Klein Center at Harvard University monitors web based media to produce interactive visualizations on geographical coverage of a news topic, important keywords corresponding to the news topic, etc. An et al. developed a platform called Media Explorer [5], which maps the US based news media sources along the political spectrum, using the co-subscription relationships inferred by Twitter links. Unlike the aforementioned platforms, our work analyzes biases at the policy level w.r.t. the individual aspects discussed within a policy issue, the five dominant frames of news presentation, and the political alignment of news-sources. Besides, we also see how these biases are carried forward by the community of Twitter followers of these news-sources. We have developed a platform which can aid in self-regulation of mass media houses in India based on this analysis.

3 DATA

To carry out our analysis of bias in mass media and social media, we have built a corpus of mass media data collected from the websites of seven leading news dailies in India, with varying commonly perceived slants towards the five constituencies proposed. Our data consists of 4 million news articles from 2011 to present, gathered on a daily basis from the following news-sources: *The Hindu, The Times of India (TOI), Indian Express (IE), The New Indian Express (NIE), Telegraph(TeleG), Deccan Herald (DecH) and Hindustan Times (HT)*, and online archives of the news-sources were used to build a corpus of news articles since 2011. We perform our analysis on 17849 articles on Demonetization (Nov 2016 to Jan 2017), 12809 articles on Aadhar (2011 to 2017), 15756 articles on GST (Jan 2011 to June 2017), and 13840 articles of Farmers' Protest (Nov 2016 to Apr 2018). Along with storing the article text, we also extract entities from each article using the *Open Calais* tool, and implement a detailed entity-resolution (ER) process to match identical entities with each other. The details of the ER process are discussed in one of our recently published works [1].

Social media data is obtained for each policy event by extracting the last 3000 tweets of every follower of the official Twitter handles of the news-sources. The number of followers corresponding to each news-source are: TOI (11026374), HT (6299716), Hindu (4842234), IE (2742132), NIE (347148), TeleG (51884), and DecH (24896). The number of tweets for the events are 396499 for Aadhaar, 1236500 for Demonetization, 1147154 for GST, and 512457 for Farmers' Protest. We refer to this set of tweets as *TweetFol* throughout the paper.

4 METHODOLOGY

The following sections describe the steps of our analysis.

4.1 Article, tweet and entity extraction

We extract articles from the media database on a given policy event based on a set of manually selected keywords related to the event as shown in table 1, which are later augmented with newer keywords from the extracted articles based on their *term frequency-inverse document frequency* (TF-IDF) scores. Similar methods based on TF-IDF have been used for keyword extraction in several studies [29–31].

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

Table 1. List of manually collected keywords used to extract articles (and tweets) corresponding to policy events

We collect tweets and retweets of all followers of a news-source handle, and extract the tweets corresponding to the policy events using the same set of keywords. We do this by first obtaining a list of all followers for every news-source, and then extracting their latest 3000 tweets (upper limit of tweets provided by Twitter for each follower) using the Tweepy API. Finally, these tweets are segregated into different policy events using the same set of keywords used to extract articles for each event. We then use the *Open Calais* service to extract entities from the news articles, which include politicians, political parties, bureaucrats, firms, managers of firms, judiciary members, and locations.

In order to keep only frequently occurring entities, we filter out all entities that occur in less than three news-sources. For each of these events, we also perform entity resolution (ER) [1] to resolve multiple entity names and aliases of the same entity occurring in different forms, to a single standard entity name. We keep a set of entities that have been successfully resolved so far, and keep augmenting it as crawling more news articles throws up additional entities to be resolved. On encountering an unresolved entity during crawling, ER within the media data follows two steps: (a) It finds the top ten candidate entities from the resolved set based on partial matching of their standard names, aliases, and context (b) It further filters these top ten entities to obtain a set of best matching entities, using string matching and phonetics based distance measures applied on standard names and context of entities. The filtering is done on experimentally set similarity thresholds [1]. The context attributes used for ER include type of entity and its standard name. Apart from these attributes, it also returns locational coordinates, state, and country information for cities. We merge this context information together, for entities that are successfully resolved with each other. This improves the ER accuracy over time as the resolver gains more and more context information for each newly resolved entity (in the course of crawling new articles). If any of these steps fail, we consider the newly encountered entity as a separate entity, and enter it separately in the resolved set. The peak performance of the ER heuristic for resolution within media data is 97.61% precision and 96.47% recall for *person* entities, and 93.82% precision and 96.2% recall for *non-person* entities.

4.2 Aspect extraction using LDA

We use Latent Dirichlet Allocation (LDA) to identify different aspects within each event. LDA is a statistical modeling method that maps a set of documents to unobserved topics, which aids in clustering similar documents into topic clusters that can be manually examined and labeled. In our case, the documents refer to the media articles, which are mapped to different topic clusters, which we refer to as aspects henceforth. Our approach in this direction is similar to [58] where the authors use LDA to cluster news topics to various aspects using both news articles on the topics and the user comments on them. We then map these LDA topics to 16 aspects for *Demonetization*, 14 aspects for *Farmers' Protest*, 11 aspects for *GST*, and 17 aspects for *Aadhar* by merging some topics together, and labeling each aspect manually. Articles are mapped to aspects for which LDA gives a probability of greater than 0.3. We used the best performing topic coherence measure as suggested in the paper by Roder et al. [45], in conjunction with the PyLDAVis package [50], to infer the optimal number of topic clusters to be specified for each policy event. To measure the accuracy of LDA aspect mapping, we manually studied 800 articles in total across all the events. Two of the authors randomly selected 200 articles from each event and assigned aspect names for each of those articles (from the list of aspects labeled after performing LDA) by reading the article text and coming to an agreement. We then check if the two aspects assigned for each article – the

[1] We use a combination of Jaro-Winkler similarity and Levenstein distance, along with substring and abbreviation matching for this step. The value of the thresholds were found to be between 0.8 to 0.9 in our experiments.

manually assigned aspect and the aspect assigned by LDA – match. We then measure the accuracy of mapping as the number of LDA mappings (for the set of 200 articles) that match the manual mappings, out of the total number of mappings. We find that the accuracies of mapping are 85% for Demonetization, 96% for Aadhaar, 81% for GSTand 76% for Farmers' Protest.

Similarly, we need to map the tweets to URLs. There have been studies on extracting sub-topics from the comments related to a textual query on social media [59]. However, in our case, we map only the tweets that contain URLs of mass media articles, to the aspects to which these articles belong, since tweets are concise and sometimes even grammatically incorrect, which makes it difficult to map them to specific aspects. The disadvantage of this approach is that the tweets which do not contain links to articles cannot be mapped to any aspect. We are looking into techniques to achieve a better tweet to aspect mapping heuristic currently. The number of tweets containing article URLs of the four events considered are 34521 for Aadhaar, 59489 for Demonetization, 38073 for GST, and 22820 for Farmers' Protest.

4.3 Sentiment and polarity analysis of articles

We experimented with the different sentiment analysis tools provided in the iFeel framework [43], and finally settled with Sentistrength for sentiment analysis of articles, and Vader [22] for sentiment analysis of tweets (since Vader specifically was designed for analysis of sentiment for short text) based on their performances in terms of accuracy. Sentistrength [53] reports TPOS (positivity) and TNEG (negativity) scores for each article. TPOS score is in the range of 1 (not positive) to 5 (extremely positive), and TNEG score is in the range of -1 (not negative) to -5 (extremely negative). The aggregate sentiment for an article is calculated as the sum of TPOS and TNEG.

To measure the accuracy of sentiment given by Sentistrength, for the 200 articles selected in the previous section to measure the performance of LDA mapping, the authors also assigned a sentiment score manually (ground truth). Articles' sentiment alignment was found to be 84% for Demonetization, 76% for Aadhaar, 60% for GST and 84% for Farmers' Protest.

4.4 Mapping of aspects to constituencies

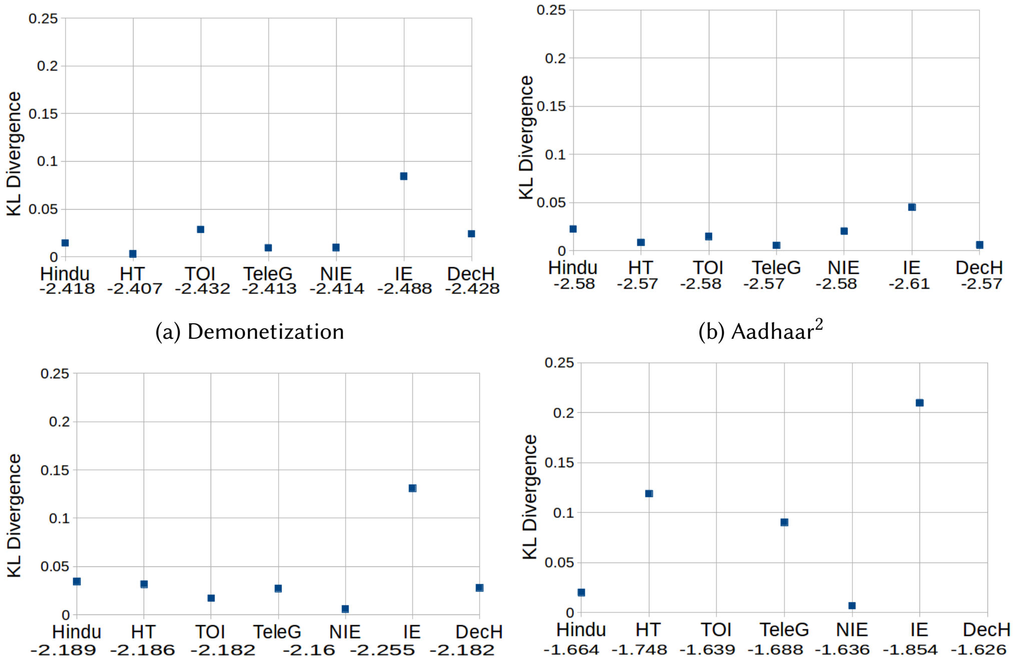
One of the goals of our work is to identify the alignment of a news-source in terms of some standard constituencies, in order to study the framing effect. We identify five constituencies: *poor*, to provide for the poor typically through wealth distribution strategies; *middle class*, typically the middle class consumers who have disposable income, to benefit through tax breaks, lower prices, and use of technology; *corporate*, driven by big corporates and formalization, economic growth, free-market policies, minimum governance; *informal sector*, driven by small enterprises and aided by slow formalization of industries and trade including agriculture; and *government* in terms of pro/anti viewpoints towards the state. For each policy event, we map each aspect to these five constituencies based on whether the aspect supports or opposes or is not applicable to the particular constituency. For example, for the *Demonetization* policy, the aspect on *Queues at banks and ATMs* is classified as pro-middle class because most articles on this aspect were negatively writing about the problems caused to the common people in getting cash at ATMs. The same aspect is classified as anti-government because negative articles on these aspects generally criticize the government's apathy and lack of foresightedness in handling the issue. These five constituencies and the news-sources' alignment to them help us in studying the effect of framing in mass media. The aspect to constituency mapping was performed by three annotators, each of whom went through around 3000 articles in total (50 articles from each aspect for each event). A coding scheme was developed to do this mapping, and is explained in the Appendix. We evaluated the inter-coder agreement for the coding exercise using the percentage agreement calculation method as described in [51]. For each policy event, we consider each (aspect,constituency) combination as a sample point, which is rated -1/0/+1 by three annotators. Based on the coding scheme, the initial mapping exercise had an inter-coder agreement of 61.33% for Demonetization,76% for Aadhaar, 71% for GST, 74.3% for Farmers' Protests. We ran another round of moderation and due deliberation before finally coming up with the final coding scheme and the current mappings. Further details of the mapping exercise are present in the *Results* section.

It is important to note that other kinds of constituencies can be added by identifying different frames. We chose the aforementioned frames because the specific economic policies considered in the paper are strongly related to these frames of content presentation and perception.

5 RESULTS

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

5.1 Ideological slant of news-sources



(a) Demonetization

(b) Aadhaar[2]

(c) GST

(d) Farmers Protest

Fig. 1. [RQ2] KL divergence across aspects between distributions of relative coverage and mean relative coverage (across news-sources) for the four policy events. Higher the deviation for a particular news source, more different is its coverage of aspects from the mean behavior across news-sources. Values on X- axis represent $\sum p * \log m$, where p represents the probablility distribution of aspects for a news-source, and m represents the mean distribution of aspects across news sources.

(RQ-1): Are news-sources biased on the amount of coverage they give to different aspects about the policy issues? This research question relates to the *agenda setting* effect of mass media as reported in literature [47]. Agenda setting is the idea that there exists a strong correlation between the emphasis placed by mass media on certain issues, and the importance attributed by the readers to these issues. In this research question, we study the emphasis that mass media places on various aspects in terms of the relative coverage given to these aspects. We define relative aspect coverage for each news-source as follows:

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

where count(words,aspect) is the number of words appearing in articles belonging to the aspect for a particular news source. When we look at the top aspects covered by mass media corresponding to each policy event in table 2, we find that for Aadhaar, the highest covered aspects generally relate to the middle class (example: *Aadhaar enrollment centers* and *Court cases related to Aadhaar*) or to the effect of the policy on economy (*Positive effect on economy*). Demonetization as a policy involved significant political debates and the highest covered aspects are all political in nature (example: *Appreciation by PM and Opposition's statements against Demonetization*) followed again by an aspect related to the economy (*Negative impact of Demonetization on economy*). GST, which is an economic policy issue, has the highest coverage provided to two aspects namely, *Political push to include petroleum products under GST*, and *Non-conclusive discussion between centre and states*. The first aspect here involved a lot of political discussion, but is also relevant to the middle class. The second aspect is a politically polar topic of discussion, which includes discussions on the losses incurred by the states in case GST is implemented. We thus observe that bias does exist in coverage provided to different aspects for each policy event, and the highest covered aspects are generally political or relevant to the middle class. However, although there is some discussion on the immediate issues of the traders and companies (*Confusions regarding GST rate slabs* and *Fears of capital crunch among traders*), the problems of the consumers – like rise in prices of essential commodities and services due to GST implementation – is not given significant attention. Finally, in case of Farmers' Protest, we find that the top covered aspects mostly focus on quick remedies to the problems of farmers (*Disbursement of loans* and *Loan waiver implementation by state governments*). There also is some discussion on structural issues related to the farmers' distress like *Irrigation concerns and water pollution.*

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

To see if the differences that seem to exist in the relative coverage given to various aspects by the news-sources are significant, we find the global relative aspect coverage for each aspect, by averaging the relative aspect coverage across all news-sources. To then see which news-sources deviate the most from the global relative aspect coverage, we plot the 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 in figure 1. We find from these plots that the highest deviation from mean aspect coverage is generally shown by IE, which is a commonly believed pro-opposition news-source [32]. TeleG and HT also show a high deviation in the case of Farmers' Protest. We also find that apart from IE, most news-sources lie close to the mean aspect coverage trend for Demonetization, Aadhaar, and GST. This reflects that for these three policies, most news-sources preferred to stay close to the average trend of aspect coverage in mass media. When we investigate the event Farmers' Protest, we find that for all news-sources the top three aspects in terms of coverage remain nearly the same. These are: *Irrigation concerns and water pollution affecting farming, Disbursement of loans and subsidies by banks for farmers, and*

Table 2. Relative aspect coverage for mass media and social media, for the top five highest covered aspects in mass media

Aspects (Aadhaar)	Mass Media	TweetFol
Aadhar enrollment centers	9.9%	8.9%
Court cases related to Aadhaar	9.9%	13.5%
International Linkages/Positive Effect on Economy	8.8%	13.6%
Implementation of Direct Benefit Transfer Scheme	6.9%	7.6%
Parliamentary debates on Aadhaar	6.3%	7.5%

(a) Aadhaar

Aspects (Demonetization)	Mass Media	TweetFol
Appreciation by PM for supporting Demonitisation	15.9%	26%
Opposition unites against government on Demonetisation	14.2%	12.6%
Negative imapct of demonetisation on Economy	8.3%	10.2%
Probes and Arrests of black money hoarders	6.1%	6.8%
Long queues at banks and ATMs and cash crunch	5.9%	4.1%

(b) Demonetization

Aspects (GST)	Mass Media	TweetFol
Protest to include petroleum products under GST	16%	16.9%
Non-conclusive discussion between centre and state	15.2%	33.1%
Discussion in the Parliament in support of GST	13.4%	3.9%
Revolts/Confusion with GST Rate Slabs	12.8%	18.9%
Fears of capital crunch among traders before GST rollout	12.3%	4.4%

(c) GST

Aspects (Farmers' Protest	Mass Media	TweetFol
Disbursement of Loans and Subsidy by Banks for farmers	12.2 %	10.8 %
Irrigation concerns and water pollution affecting farming	9.7 %	22.8 %
Loan waiver implementation by State Govts	7.5 %	3%
Protests by farmers	7.1%	6.1%
Farmers' Distress regarding Minimum Support Price for Crops	5.6%	2.8%

(d) Farmers' Protest

Protests by farmers. However, while most of the other news-sources provide a significant coverage to other aspects too, IE provides most of its coverage to these three aspects, and negligible coverage to others. Owing to this concentration of coverage to the top three issues, IE shows a high deviation from the mean aspect coverage trend.

(RQ-2): 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 analyze the effect of *framing* in mass media [47]. Framing refers to the modes of presentation that media houses use 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 analyze this effect by automatically extracting aspects from the news articles, and manually linking them with one or more of these five constituencies based on the coding schemes designed for each policy event. As mentioned earlier, this mapping simply tells us if articles in that aspect contain keywords semantically similar to the constituency name, or if the articles discuss about issues pertinent to the constituency. Next, we find the alignment of the aspect towards the constituencies in terms of the sentiment of its constituent articles. This is done in two steps: (a) we first find out the majority sentiment slant *m* of the articles in an aspect *a*, and (b) we then see if the majority articles of the aspect indicates whether it supports or is against a particular constituency *c*, i.e., its stance w.r.t. the constituency or *stance(a,c)*. We divide *stance(a,c) by the majority sentiment m, to get the alignment score (U) of the aspect w.r.t. the constituency. U* can vary between -1, 0, and +1 for each (aspect, constituency) pair. These scores are presented in the Appendix. Using these (aspect,constituency) alignment matrices for the four events, we calculate the (news-source,constituency) alignment matrix *M* as follows:

$$C(i, a, n) = \frac{count\,(words, i)}{\sum_{j \epsilon (n, a)} count(words, j)} \tag{1}$$

$$S(n, a) = \sum_{i \epsilon (n, a)} C(i, a, n) * (S(i, a, n) - S_{avg}(a)) \tag{2}$$

$$M(n, c) = \sum_{a \epsilon c} U(a, c) * S(n, a) \tag{3}$$

where *n* represents a news-source, *a* an aspect, $C(i, a, n)$ is the relative coverage for the ith article, in news-source *n*, belonging to aspect *a*, and $S(i, a, n)$ is the compound sentiment score of the ith article for aspect *a*. $S_{avg}(a)$ is the average sentiment score of all articles in aspect *a* across all news-sources, and *c* is the constituency. Here, $(S(i, a, n) - S_{avg}(a))$ is the offset of the sentiment of article-i from the mean sentiment of all articles for aspect a across news-sources, $U(a, c)$ is the (aspect,constituency) alignment value $\epsilon[-1, 0, +1]$. Thus, the matrix *M* tells us how aligned a news-source is to the aggregate behavior of all news-sources, for a constituency, in terms of the coverage and sentiment deviation with which it presents its content. To empirically verify if there exists variations (or similarities) in terms of constituency alignment of the news-sources, we performed a Principal Component Analysis (PCA) on the 5-dimensional mean constituency vector (mean of the 5-dimensional constituency vectors across all events) for each news-source, for the four events. Figure 2 shows the plot. A factor analysis was then done to interpret the two principal components we obtained here.

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

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

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

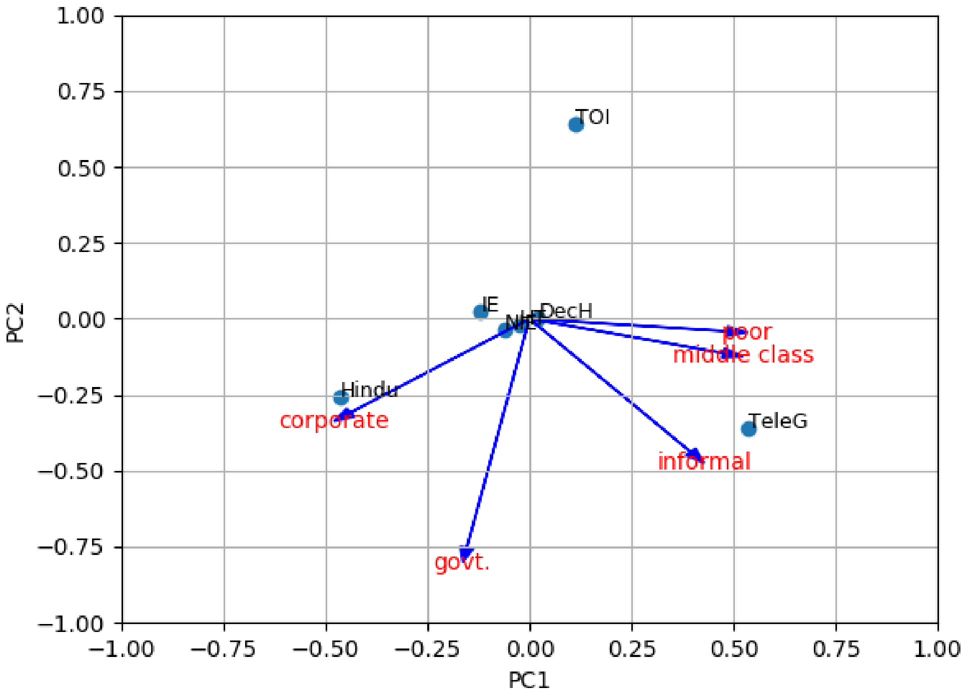


Fig. 2. PCA on constituency vectors for the four events: Principal component PC1 represents news-sources that cover more of informal sector, poor, and middle class (towards right) related issues, and political or corporate related issues (towards left). Principal component PC2 represents news-sources that cover political, corporate, and informal sector related issues (on the negative side).

coverage to issues related to the poor, and more coverage to middle class and political issues. We present these findings in the Appendix.

The PCA analysis and the analysis of the mean relative coverage provided to constituencies by mass media indicates the presence of *framing* effects of mass media: (a) the news-sources are biased w.r.t. the five constituencies, in terms of the coverage and sentiment with which they present their content, and (b) they provide consistently less coverage to issues of the poor in general.

5.2 Alignment of news-sources with their readership in social media

In RQ1, we observed that the news-sources are biased on the amount of coverage that they provide to different aspects belonging to policies. In continuation of that, here we analyze if the readers' preferences in terms of the importance placed on certain aspects, and the sentiment slants of their posts, correlate with that of mass media. We consider the readership community of news-sources as the set of all followers of the news-source handles on Twitter (*TweetFol*). We next attempt to answer the following research questions:

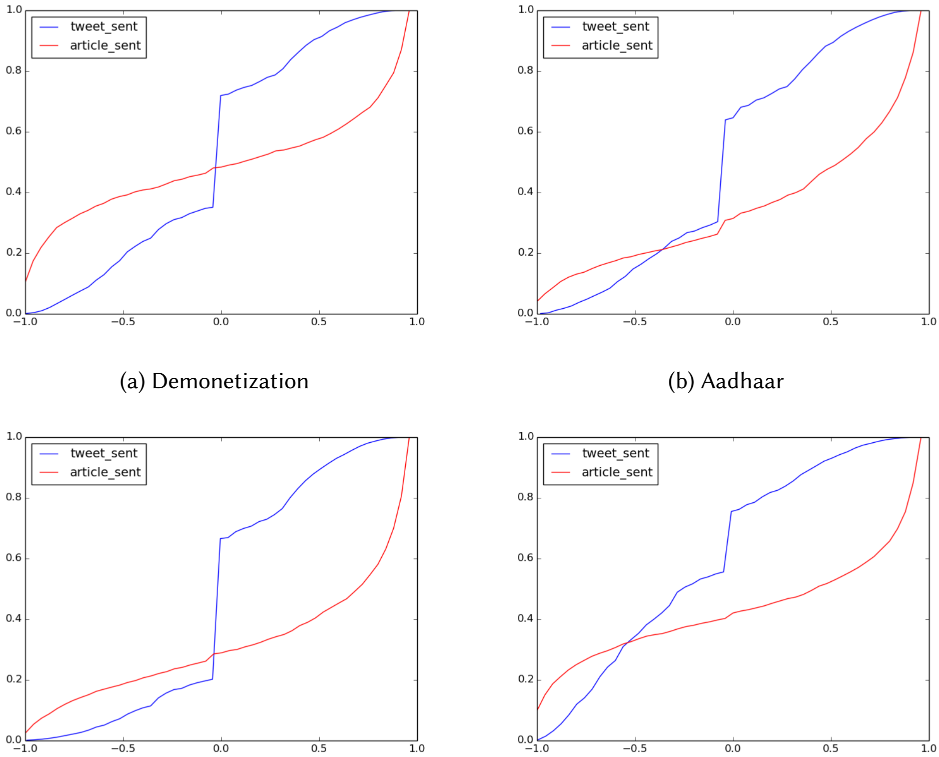
(RQ-3:) Are some news-sources more closely aligned with their readers (on social media) than others? Under this research question, we examine whether the comments made by the readers of a news-source on Twitter (in terms of the aspects posted about, and the sentiment slant) align with that of the news source itself. We see if the aspects tweeted by the readers of news-sources align with the ones presented by mass media. For each news-source, we compute the Jensen-Shannon Divergence (JSD) between the distribution of its aspect coverage and that of its social media community. The Jensen-Shannon divergence is a principled divergence measure that quantifies how distinguishable two or more distributions are from each other [2]. Table 3 depicts our findings. As we can observe, for each policy, the news-sources have a high alignment in terms of aspect coverage with their followers on social media (as seen from the low values of JS divergence), and the readers of the news-sources prefer to closely follow the aspect coverage trend of their favorite media houses. We are able to see, therefore, that both the mass media and the social media give less coverage to issues of the poor, strongly indicating a bias arising in the web content produced in the mass media and social media, similar to the biases caused due to the digital divide. Digital divide is defined as an uneven distribution in the access to, use of, or impact of information and communication technologies (ICT) between any number of distinct groups, which may be defined based on social, geographical, or geopolitical criteria, or otherwise [16, 36]. It must be noted that our observations do not indicate that the digital divide causes these biases, but that similar biases arise due to the digital divide. In our case, we see these biases arising due to the alignment in agenda or constituency covered in the content produced by the mass media and social media.

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

Table 3. [RQ3] JS divergence showing difference in aspect coverage between mass media and social media: for TeleG, we could not find any tweet for Demonetization and GST

We next analyze for each news source, how much the overall sentiment of the Twitter posts on its news articles align with the sentiment of the original article. In figure 3, we show the CDFs for article sentiment and tweet sentiment (note that in this case, we consider all tweets by the followers of a news-source, irrespective of whether they contain URLs or not). The plots indicate two interesting observations: (a) more than 20% of the tweets by followers of news handles are neutral (in Demonetization, Aadhaar, and GST more than 40% of the tweets are neutral), and (b) the article sentiment is either more positive, or more negative than the tweet sentiment (as observed from the curves to the right and left of the neutral axis). Both of these trends indicate that in general, the Twitter followers of news-source handles are more neutral than the news-sources themselves[3]. We present the results for the other news-sources in the Appendix, which are consistent with the

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



(a) Demonetization

(b) Aadhaar

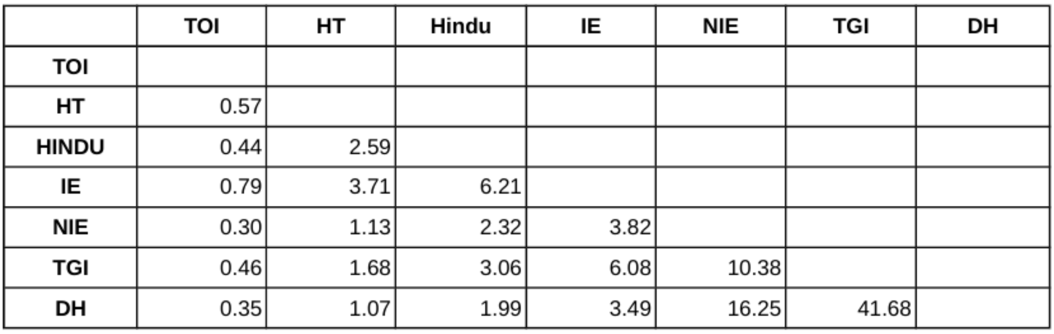
(c) GST

(d) Farmers’ Protest

Fig. 3. CDF plot of article sentiment and tweet sentiment for the set TweetFol, for *The Hindu*. For the other news-sources for all events, we present the results in the Appendix.

observations for Hindu. These findings tell us that social media seems to respond to whatever is being discussed in the mass media, but the sentiments of the social media readership are not entirely aligned with the news-sources that they follow.

Considering this high degree of alignment of aspect coverage between the mass media houses and their Twitter community, we next want to check whether these communities are distinct from each other. We therefore analyze the community overlap between the followers of different news-sources and present our results in table 4. We broadly find that communities of DecH, TeleG, NIE, IE, and Hindu form a closely knit cluster in terms of their community overlap. Among these news-sources, DecH, TeleG, and NIE show the highest overlap of communities amongst themselves (odds ratio for community overlap > 10). HT and TOI have least overlap with others, and form outliers (odds ratio for community overlap < 4). This indicates that many of the followers prefer to follow news-sources with commonly believed ideological affiliations that are opposite to each other (DecH is commonly believed to be pro-opposition and NIE is commonly believed to be pro-ruling party). This might be an indication of users’ tendency to consume news from news-sources with opposing polarities, to counter the information bias in media. Our findings partly support the findings from the study by Mullainathan et al. [33], which models the mass media to see the effect of competition among media houses on two types of biases – ideological bias (bias towards or against



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

Table 4. Odds-ratio of overlap of follower community for each pair of news-sources

a political ideology), and spin bias (bias occurring due to propensity of media towards creating a memorable story). They observe that although competition can aid in removing ideological biases of media, it exaggerates the incentive of spinning stories. Our analysis provides a direction to study the open question of whether the newspapers tune their coverages to bring them closer to what their follower communities want, or whether communities of people gravitate to the news-sources that align with their ideologies, or both. It raises the question of what other factors come into play for social media followers to decide which news-sources to follow? This decision function seems to not only be dependent on the biases in the content, but may have more factors included, such as differences in the popularity of the newspapers, and pre-conceived preferences of the users. This is an open question, and can be investigated further to understand the rationale behind the choice of which news-sources to follow.

6 DISCUSSION AND CONCLUSION

We discuss in this section, the broad findings of our analysis in terms of the three research questions that we study:

Ideological slant: We find that variation exists in the coverage of different aspects across news-sources. Our analysis on mean coverage provided to each constituency also suggests that the news-sources generally provide high coverage to political issues and issues related to the middle class. On the other hand, the issues of the poor do not get enough attention in comparison. The PCA analysis based on coverage and sentiment of the content also indicate biases in the alignment towards the five constituencies.

Alignment of news-sources with audience on social media: We also find that social media is more balanced in taking up the views of academics and activists for discussion and distribution; politicians still get the most coverage but less than that given by mass media. The aspects covered by the readers of news-sources is closely aligned with the those covered by the news-sources themselves. However, the readers (in this case, followers of the Twitter handles of the news-sources) are slightly more neutral in their tweets compared to the news-sources whose URLs they post. This tells us that the readers of the news-sources have similar slants towards constituencies and interest in aspects as covered by the news-sources they follow, although they do show some independence in their personal opinions in terms of the sentiment slant.

These findings about biases raise questions on the role of media. Siebert defines the four theories of press [49] namely, the *Authoritarian*, the *Soviet Communist*, the *Libertarian*, the *Social Responsibility* theories. According to the Authoritarian theory, the press and all of the information contained in it is controlled by the state or the government. The Soviet Communist theory provides higher control of the state on the press or media. According to this theory, not only does the government control the media or the information present in the media, but it also runs the media as a tool for its own propaganda. Libertarian theory stands opposite to the Authoritarian theory, and keeps the press or media out of state control. According to this theory, the primary duty of the press is to serve the interest of citizens by presenting the truth. Since a capitalistic society allows for free enterprises and corporate control of press or media, Social Responsibility theory states that the press should be made strong enough to function outside of any influence, be it corporate influence or state influence. We find from the last two theories that the role of media is to analyze and critique the state's policies, acting as a watch dog. Moreover, apart from informing, entertaining, and acting as a watchdog, media also has the responsibility of raising conflicts to the plane of discussion. With mass media and social media being one of the significant contributors of web data, the presence of biases suggests a stronger need for self-regulation of the Indian mass media, with respect to these pivotal roles. Further, the evident lack of coverage provided to the immediate issues of the poor enlarges the problem of digital divide. Our work serves as an initial step to address this issue by providing empirical justification of less representation to the issues of the poor, who might not have access to digital technology or social media. Moreover, in order to ensure that the algorithms suggesting the news items to users do not reproduce these biases, we are currently building a recommendation system, which ensures long term fairness and short term diversity in representation of the various constituencies to which the news belongs. We believe that this platform and the techniques used can bring more visibility to the functioning of mass media, and push it closer to the goal of achieving diversity in content publication and educating the public of different viewpoints.

REFERENCES

[1] [n. d.]. Blinded for review. ([n. d.]).

[2] 2019. Jensen-Shannon Divergence. (2019). https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence

[3] 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, 36–43. https://doi.org/10.1145/1134271.1134277

[4] Aayush Ailawadi. 2017. Your Aadhaar Data Is Now With Private Companies As Well. (2017). https://www.thequint.com/news/business/aadhar-data-with-private-companies

[5] Jisun An, Meeyoung Cha, Krishna Gummadi, Jon Crowcroft, and Daniele Quercia. 2012. Visualizing media bias through Twitter. In *Sixth International AAAI Conference on Weblogs and Social Media*.

[6] 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*.

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

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

[9] Ricardo Baeza-Yates. June 2018. Bias on the Web. *Commun. ACM* 61, 6 (June 2018), 54–61.

[10] Ricardo Baeza-Yates, Carlos Castillo, and Efthimis N. Efthimiadis. 2007. Characterization of National Web Domains. *ACM Trans. Internet Technol.* 7, 2, Article 9 (May 2007). https://doi.org/10.1145/1239971.1239973

[11] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[12] 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), 250–271.

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

[14] Simeon Djankov, Caralee McLiesh, Tatiana Nenova, and Andrei Shleifer. 2001. *Who owns the media?* Technical Report. National Bureau of Economic Research.

[15] Gillian Doyle. 2002. *Media ownership: The economics and politics of convergence and concentration in the UK and European media*. Sage.

[16] Wikipedia The Free Encyclopedia. [n. d.]. Digital Divide. ([n. d.]). https://en.wikipedia.org/wiki/Digital_divide

[17] Wikipedia: The Free Encyclopedia. 2016. 2016 Indian banknote demonetisation. (2016). https://en.wikipedia.org/wiki/2016_Indian_banknote_demonetisation

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

[19] Berkman-Klein Center for Internet and Harvard University Society. [n. d.]. Media Cloud. ([n. d.]). https://mediacloud.org/

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

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

[22] 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*.

[23] E. Goldman. 2008. *Search Engine Bias and the Demise of Search Engine Utopianism*. Springer Berlin Heidelberg, Berlin, Heidelberg, 121–133. https://doi.org/10.1007/978-3-540-75829-7_8

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

[25] Jürgen Habermas. 1989. The structural transformation of the public sphere, trans. Thomas Burger. *Cambridge: MIT Press* 85 (1989), 85–92.

[26] 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.). (1988).

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

[28] Joseph T Klapper. 1960. The effects of mass communications. (1960).

[29] Sungjick Lee and Han-joon Kim. 2008. News keyword extraction for topic tracking. In *2008 Fourth International Conference on Networked Computing and Advanced Information Management*, Vol. 2. IEEE, 554–559.

[30] Juanzi Li, Kuo Zhang, et al. 2007. Keyword extraction based on tf/idf for Chinese news document. *Wuhan University Journal of Natural Sciences* 12, 5 (2007), 917–921.

[31] Yutaka Matsuo and Mitsuru Ishizuka. 2004. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools* 13, 01 (2004), 157–169.

[32] Atul Kumar Mishra. 2015. Newspapers in India and their Political ideologies. (2015). https://rightlog.in/2015/07/newspapers-in-india-and-their-political-ideologies/

[33] Sendhil Mullainathan and Andrei Shleifer. 2002. *Media bias*. Technical Report. National Bureau of Economic Research.

[34] Sendhil Mullainathan and Andrei Shleifer. 2005. The market for news. *American Economic Review* 95, 4 (2005), 1031–1053.

[35] Sevanti Ninan. 2007. *Headlines from the heartland: Reinventing the Hindi public sphere*. Sage.

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

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

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

[39] By Babatunde Oshinowo Jr. [n. d.]. Examining Bias and Distortion in Mass Media in America. ([n. d.]).

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

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

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

[43] Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5, 1 (07 Jul 2016), 23. https://doi.org/10.1140/epjds/s13688-016-0085-1

[44] 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).

[45] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*. ACM, 399–408.

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

[47] 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), 9–20.

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

[49] Fred Siebert, Theodore Bernard Peterson, Theodore Peterson, and Wilbur Schramm. 1956. *Four theories of the press: The authoritarian, libertarian, social responsibility, and Soviet communist concepts of what the press should be and do.* University of Illinois press.

[50] Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. 63–70.

[51] Stephanie. 2016. Inter-rater Reliability. (2016). https://www.statisticshowto.datasciencecentral.com/inter-rater-reliability/

[52] Newslaundry Team. [n. d.]. Who owns your media? http://www.newslaundry.com/2014/02/05/who-owns-your-media-4/. ([n. d.]). Accessed: 17/4/2016.

[53] 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), 2544–2558. https://doi.org/10.1002/asi.v61:12

[54] Liwen Vaughan and Mike Thelwall. 2004. Search engine coverage bias: evidence and possible causes. *Information processing & management* 40, 4 (2004), 693–707.

[55] Claudia Wagner, David García, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *CoRR* abs/1501.06307 (2015). arXiv:1501.06307 http://arxiv.org/abs/1501.06307

[56] Wikipedia. Updated on March 2018. Kisan Long March, Maharashtra. (Updated on March 2018). https://en.wikipedia.org/wiki/Kisan_Long_March,_Maharashtra

[57] Wikipedia. Updated on May 2018. Goods and Services Tax (India). (Updated on May 2018). https://en.wikipedia.org/wiki/Goods_and_Services_Tax_(India)

[58] Sevgi Yigit-Sert, Ismail Sengor Altingovde, and Özgür Ulusoy. 2016. Towards detecting media bias by utilizing user comments. In *Proceedings of the 8th ACM Conference on Web Science*. ACM, 374–375.

[59] Hang Zhang and Vinay Setty. 2016. Finding diverse needles in a haystack of comments: social media exploration for news. In *Proceedings of the 8th ACM conference on web science*. ACM, 286–290.

Appendices

A CODING SCHEMA

We describe the coding schema taking Demonetization as an example, in this section. The coding schemes for all of the other policy events are created similarly. The goal of the scheme is to determine whether articles of some aspect talk about a particular constituency or not. For example, articles that refer to labourers who do manual work with daily wages in the construction sector, are considered to be talking about the poor constituency, and articles that talk about workers in the IT industry, are considered to be talking about the middle class. To ensure consistency across different events, these definitions are suitably customized for each event. Thus, for farmer protests, articles referring to smallholder farmers and tribals are considered as discussing the poor, while articles discussing urban consumers are considered to be talking about the middle class. This helps the annotator understand the general definition of the constituency based on that policy event. The coding schema was built after sampling roughly 100 articles for each policy event, by the lead author of this study and assisted by two annotators. The annotators then went to perform the final aspect to constituency mapping. After finalization, the coding scheme was given to the annotators to perform the final aspect to constituency mapping.

Coding Schema for Demonetization

Constituency	Article primarily targets:	Examples	Normative Definition
Poor	- Labourers, workers in factories and small mills (e.g., textile and diamond-cutting mills), migrant workers and labourers, poor people belonging to the lowest level of income, and workers without bank accounts - Welfare schemes like PMGKDS (recovering black money to help the poor) and discussion about their success or failures - Railways, passengers, tickets of public modes of transport (buses, trains, etc.) availed by lower income groups - Retail inflation, inflation, consumers, consumer price index (CPI), consumer food price index (CFPI), names of food items. - Black money hoarding related articles with reference to benefits caused to poor	The city at present has 9000 manufacturing units which employ 7 lakh workers. In November, by the time the demonetization was announced, at least some wages were paid and some were managed during the last two weeks, but now with December around the corner, how are we going to pay the wages if we do not have a cash flow. Rs 50,000 per week is not adequate for a factory as each employ over 2000 labourers.	Poor people at the lowest levels of income. This includes labourers and factory workers without bank accounts. The welfare schemes which target the poor directly, like PMGKDS also come in the ambit of this class. Casual workers (working on contractual basis) with daily wage below 200 INR.

Middle class	- Workers employed in sectors with higher income range (e.g., daily wagers working in garment based activities like stitching), workers for whom absence of bank accounts is not specifically mentioned. - ATMs, cash withdrawal limit - Note exchange at post offices, banks, customers, - Long queues - City residents and office workers belonging to the service sector (e.g., working at public offices, MNCs, etc.). - Black money hoarding related articles (preferably with reference to the middle class), income tax, tax evasion. -Retail inflation, inflation, consumers, consumer price index (CPI), consumer food price index (CFPI), names of food items. - Railways, passengers, tickets of flights/trains/buses, public transport system (buses, trains, autorickshaws*). - E-POS (point of sales system)	The currency stock at banks across the city has improved but most ATMs still display the message temporarily unable to dispense cash. A large number of people who thronged ATMs to take advantage of the increased withdrawal limit were left disappointed on Tuesday as majority of the machines were empty.	Middle class people who suffered the immediate aftermath of the policy move like standing in long queues at ATMs, lack of money exchange at banks and post offices, and so on. Workers and daily wagers mentioned in general (absence of bank accounts not specifically mentioned) come in this class. Passengers of public transport. Regular workers/daily wage earners (employed on a permanent basis) with daily wage above 200 INR.
Corporate	- Manufacturing companies, industries, MSMEs, factories, multinationals, businesses, big real estate companies - Entrepreneurs, businessmen, bizmen - Import, export, raw material - brands, marketing - Sensex, investors, NSE, NIFTY, BSE, foreign capital,	The ban on currency notes has brought business in the industrial city almost to a halt. Due to cash crunch , it has become difficult for industrialists to buy raw material, pay bills and make payments to labourers.	Big business houses, industrialists, SMEs and MNCs, and corporate business houses in general.

Informal sector and small traders	- Unorganized sector, informal sector, companies not registered, unregistered enterprises - Small vendors/businesses, garment sellers, paanwallahs, chaiwallahs, shopkeepers, vegetable sellers, hawkers, barbers, autorickhshaw drivers, sweet sellers, grocery shops, roadside vendors	The impact of demonetisation on ancillary units in the unorganised sector is likely to have a cascading impact on registered manufacturing units in the long run.While there are no accurate records of the number of such units in Mysuru since they are not registered and hence officially illegal they are the backbone of medium and small scale units in the region as they provide services at throwaway rates.	Unorganized sector, unregistered companies, small traders, and vendors.
Government	- State/Central government, state, centre - Name of prominent politician, minister, ministry, MP/MLA, their relatives - Names/positions of important government officials and designations (like CBDT, principal director income tax, etc.) - Failure of welfare schemes like Jan Dhan accounts - Charges of corruption against politicians and their relatives	Additionally, after a huge surge in deposits, Jan Dhan accounts witnessed net withdrawal of Rs 3,285 crore in the last fortnight. This was despite the fact that the monthly upper withdrawal limit was fixed at Rs 10,000 per month from November 30 to check misuse of Jan Dhan accounts.	State and central government, policy makers, ministers, ministries, MPs, MLAs, and their relatives. Discussions in Parliament or assemblies about the narrative on Demonetization also come in this class.

B ASPECT TO CONSTITUENCY ALIGNMENT MATRICES

The final aspect to constituency mappings obtained after the annotation are shown in these tables. The value (+1/-1) in each cell indicates the alignment value for that (aspect,constituency) combination, which when multiplied by the majority sentiment slant of that aspect, results in the final stance (pro/anti) of the aspect towards the constituency. An alignment value of 0 indicates that the aspect is unrelated to the constituency.

Aspect	poor	Middle Class	Corporate	Informal Sector	Government
Failure of RBI to stabilize economy and answer questions raised post demonetisation	0	-1	-1	0	1
Negative impact of Demonetisation on small and medium scale industries and its employees	-1	-1	-1	-1	1
Long queues at banks and ATMs and cash crunch	0	1	1	1	0
Court verdicts related to demonetisation and penalties issued for black money hoarders	1	1	-1	0	1
Vendors going cashless and negative impact on them due to demonetisation	0	-1	0	-1	1
Probes and arrests of black money hoarders	0	-1	1	0	-1
Suffering of farmers and relief measures for them	-1	-1	0	-1	1
Huge deposits in PMY bank accounts post demonetisation and announcement of minimum balance requirements	-1	-1	0	0	1
Negative imapct of demonetisation on rural economy,national economy,industries,GDP,job creation etc.	0	-1	-1	0	1
Appreciation by PM for supporting Demonitisation	-1	-1	-1	-1	-1

Table 5. Alignment matrix for Demonetization

Aspect	poor	Middle Class	Corporate	Informal Sector	Government
Positive effect of climatic conditions on agriculture yield	-1	0	0	0	0
Opposition's protests and concerns on problems related to farmers (including Demonetization)	1	0	0	0	-1
Educational Drives and Social Awareness on Farmer's Distress	-1	0	0	0	0
Politics in Maharashtra/Punjab over Compensation and Loan waiver for farmers	-1	0	0	0	1
Variation in Crop Prices with monsoon season	-1	-1	0	-1	0
Cultural events related to farmers' distress	1	0	0	0	0
Disbursement of Loans and Subsidy by Banks for farmers	-1	0	0	0	1
Crimes and Suicide in farmer community	-1	0	0	0	1
Protests by farmers	-1	0	0	0	0
Loan waiver implementation by State Govts (and opposition's protests regarding the same)	-1	0	0	0	1
Training programmes and seminars to improve farming in States	1	0	0	0	1
District Administration and farmers Issues	-1	0	0	0	1
Farmers' Distress regarding Minimum Support Price for Crops	-1	0	0	0	1
Irrigation concerns and water pollution affecting farming	-1	0	0	0	1

Table 6. Alignment matrix for Farmers' Protests

Aspect	poor	Middle Class	Corporate	Informal Sector	Government
Requirement of Aadhaar for passport and other services (concessions)	0	1	0	0	1
Fake ration cards caught due to Aadhaar linkage, aiding in the good of poor and middle class	1	1	0	0	1
Installation of e-pos systems for Aadhaar enabled PDS causing resentment among poor and middle class	-1	-1	0	0	1
Digitization of Aadhaar enabled employees' provident fund, attendance systems at public offices, and cashless payments helping the middle class	0	1	0	0	1
Requirement of Aadhaar for school admission and the middle class	0	1	0	0	1
Linking of Aadhaar with different schemes like PAN, mobile numbers, and bank accounts	0	1	0	0	1
Implementation of Direct Benefit Transfer Scheme	1	1	0	0	1
Aadhaar based verification for middle class telecom users and data leakage charges against telecom companies	0	1	0	0	-1
Opening of bank accounts and financial inclusion helping the poor	1	1	0	0	1
LPG Subsidy & DBT (Direct Benefit Transfer) helping the poor and middle class	1	1	0	0	1

Table 7. Alignment matrix for Aadhaar

Aspect	poor	Middle Class	Corporate	Informal Sector	Government
Sensex/Market rejoices on GST rollout	0	1	1	1	1
GST Bill implementation in State Legislative Assemblies	0	1	1	1	0
Traders hassles to meet GST registration deadline and changes in sensex/nifty	0	0	-1	-1	-1
Fears of Capital Crunch among traders before GST Rollout	0	0	1	0	1
Training Programmes for Creating GST Awareness	0	0	1	1	1
Centre-State deadlock in GST implementation	0	0	0	0	1
Uncertainity in Impact of GST on consumer Goods	-1	-1	1	0	0
Confusion amidst implementation of GST rate slabs	1	1	0	0	-1
Effect of GST on GDP and Economy	0	1	0	0	0
GST Bill Discussion in Parliament	0	0	1	0	1

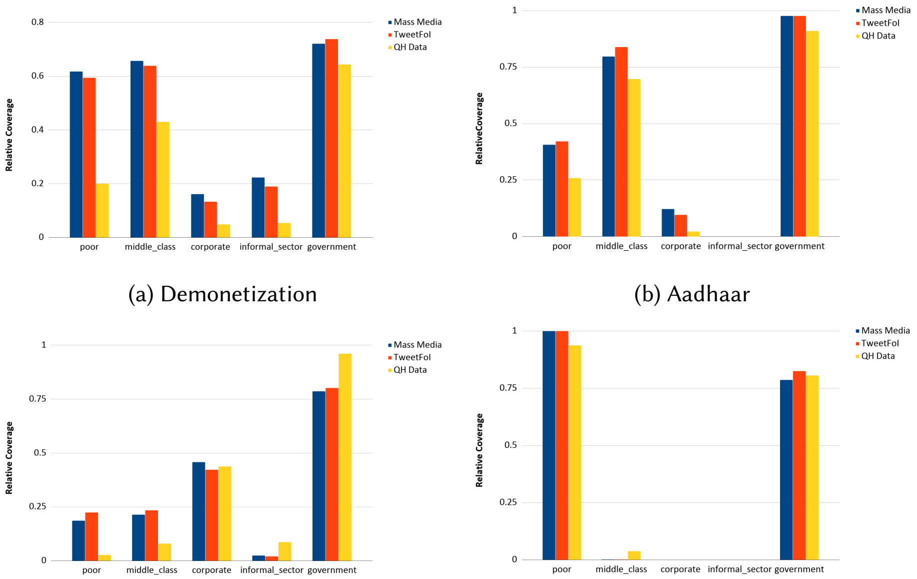
Table 8. Alignment matrix for GST

C COVERAGE OF CONSTITUENCIES BY MASS MEDIA

We analyze the relative coverage provided by mass media to the five constituencies of *poor, middle class, corporate, informal sector, and government.* The relative coverage for a constituency is calculated from the aspects belonging to the constituency, for all news-sources, across all events.

$$Cov(const) = \frac{\sum_{a \epsilon const} count(words, a)}{\sum_{asp \epsilon A} count(words, asp)} \tag{4}$$

where *Cov(const)* is the relative coverage provided to the constituency, *a* and *asp* are aspects, *A* is the set of all aspects across all events across news-sources, and *count(words,a)* is the total number of words across all articles (across all events, across news-sources) for aspect *a*. We show the constituency coverage for each policy even in figure 4.



(a) Demonetization

(b) Aadhaar

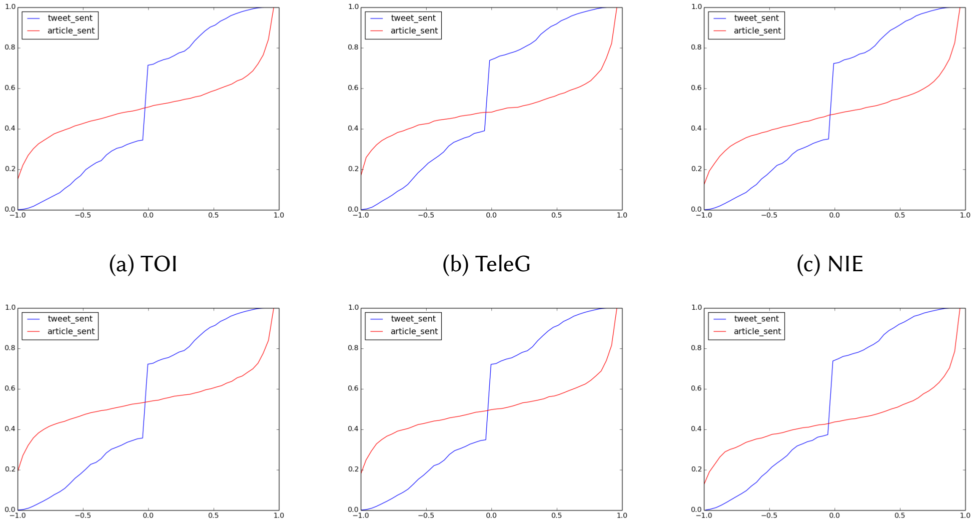
(c) GST

(d) Farmers' Protest

Fig. 4. Relative coverage provided by mass media, social media community, and QH data to each constituency for the policies

D ALIGNMENT OF NEWS-SOURCES WITH THEIR READERS

In this section, we show our results for Demonetization and Aadhaar, corresponding to the research question RQ-3: *Are some news-sources more closely aligned with their readers (on social media) than others?*. In figures 5, 6, 7, and 8, we report the news-source wise cumulative distribution functions (CDFs) for the sentiment distribution of mass media and social media, for all news-sources except *The Hindu*, which is present in our main paper (section 5.2). Our findings tally with the results that we described in the main paper.



(a) TOI

(b) TeleG

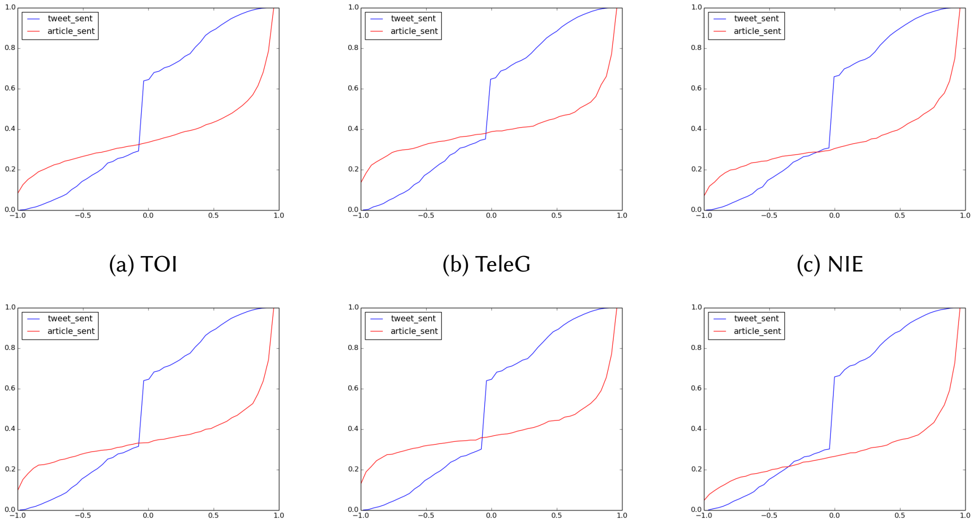
(c) NIE

(d) IE

(e) HT

(f) DecH

Fig. 5. CDF plots of article sentiment and tweet sentiment for the set TweetFol, for Demonetization, across news-sources except The Hindu. The plot for *The Hindu* is reported in the main paper.



(a) TOI

(b) TeleG

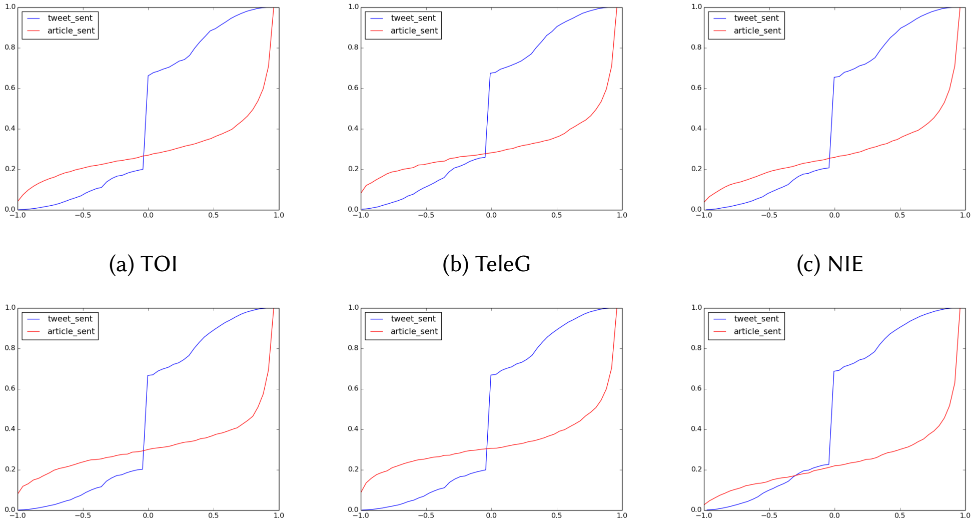
(c) NIE

(d) IE

(e) HT

(f) DecH

Fig. 6. CDF plots of article sentiment and tweet sentiment for the set TweetFol, for Aadhaar, across news-sources except The Hindu. The plot for *The Hindu* is reported in the main paper.



(a) TOI

(b) TeleG

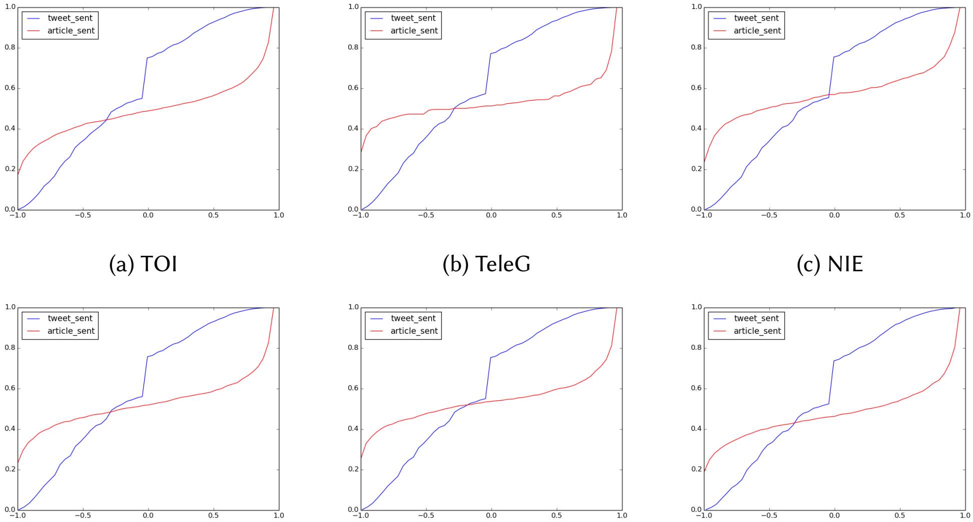
(c) NIE

(d) IE

(e) HT

(f) DecH

Fig. 7. CDF plots of article sentiment and tweet sentiment for the set TweetFol, for GST, across news-sources except The Hindu. The plot for *The Hindu* is reported in the main paper.



(a) TOI

(b) TeleG

(c) NIE

(d) IE

(e) HT

(f) DecH

Fig. 8. CDF plots of article sentiment and tweet sentiment for the set TweetFol, for Farmers' Protest, across news-sources except The Hindu. The plot for *The Hindu* is reported in the main paper.