Ideology Detection in the Indian Mass Media

Ankur Sharma* IIT Delhi ankursharma.iitd@gmail.com
Navreet Kaur* IIT Delhi navreet700@gmail.com
Anirban Sen IIT Delhi sen.anirban09@gmail.com
Aaditeshwar Seth IIT Delhi aseth@cse.iitd.ac.in

***Abstract*—Ideological biases in the mass media can shape public opinion. In this study, we aim to understand ideological bias in the Indian mass media, in terms of the coverage it provides to statements made by prominent people on key economic and technology policies. We build an end-to-end system that starts with a news article and parses it to obtain statements made by people in the article; on these statements, we apply a Recursive Neural Network based model to detect whether the statements express an ideological bias or not. The system then classifies the stance of the non-neutral statements. For economic policies, we determine if the statements express a pro or anti slant about the policy, and for technology policies, we determine if the statements are positive or skeptical about technology. The proposed research method can be applied to other domains as well and can serve as a basis to contrast social media self-expression by prominent people with how the mass media portrays them.**

***Index Terms*—Ideology Detection, Ideology Classification, Media Bias, Social Policy, Social Media Analysis, Mass Media Analysis, Sentiment Analysis, Recursive Neural Networks**

I. Introduction

Mass media can significantly shape public opinion [1], and the study of different forms of biases in the mass media is an active area of study [2, 3]. In this paper, we focus on the statements made by important people (the power elite - politicians, business leaders, economists, administrators [4]) quoted in the mass media, and demonstrate a computer-based system to identify the ideological slant of these statements. We identify slants in two domains: whether the statements reveal a pro/anti slant towards economic policies, and whether they reveal a technology deterministic/skeptical slant – technology determinism considers technology to be a solution to all problems in the society; the opposite being technology skepticism [5]. Our contribution also includes providing an open annotated dataset of 3855 pro/anti statements made by the power elite on four economic policies, and 812 statements made by them related to four technology policies, in India. Our system operates in an end-to-end manner, starting with the crawling of news articles from six mainstream English dailies, followed by entity identification of people mentioned in these articles, extraction of statements made by them, classification of these statements into neutral and ideologically biased (non-neutral) classes, and identifying the nature of bias (positive/negative and deterministic/skeptical) for the non-neutral statements. We use different machine learning tools at different steps, and in this paper we present a deeper focus on the ideology classification component developed using Recursive Neural Networks (ReNNs). We also briefly present an application of our system to understand the ideological biases of some of the most prominent politicians in India, and six prominent English national newspapers in terms of the coverage they provide to these politicians.

Our work is relevant in the field of social computation to understand ideological biases in mass media, and can augment the research of social media expressions made by the same entities on platforms like Twitter. For instance, this can reveal insights about whether the mass media editorially introduces biases over what the entities may self-express on social media.

Most studies to examine ideological biases expressed in the media have used sentiment analysis tools that were primarily developed to track product reviews [6]. These tools indicate how much of a positive or negative slant a statement has, but they fall short in capturing the ideological stance. For instance, SentiStrength [7], a popular tool for sentiment analysis, classifies the following statement *"Just because it is possible to hack a network does not mean that technology must not be deployed."* with a strong negative sentiment although for our purposes we would like it to be classified as a pro-technology statement. Ideology classification, therefore, requires us to pick up complex linguistic features such as common phrases that are used in the context of different policies, sentence structures, etc. Our ReNN model for ideology detection is intended to capture such features.

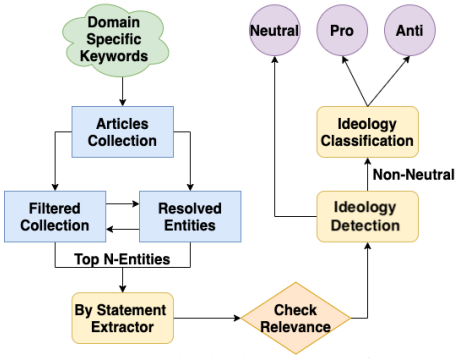


Fig. 1: Our proposed Ideology Detection Framework

The end-to-end system developed by us is shown in

*These authors contributed equally to the paper.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00

figure 1. It consists of three main components: **(a) Data extraction** to obtain statements made by various entities in different newspapers; **(b) Relevance filter** to remove statements that are not relevant to the policies of interest; and **(c) A two-step classifier** which first checks whether a statement has no stance (i.e. neutral), such as a factual statement made by an administrator about a policy, and then for non-neutral statements to check whether they are pro/anti on economic policies, or deterministic/skeptical towards technology policies.

We took care in building the ideology classifier to ensure that it does not classify based on the entity, but rather on what the entity stated. Ruling party politicians who bring in new policies can otherwise be expected to talk positively, while opposition party politicians would speak negatively. We, therefore, blinded out all entity mentions in our dataset to avoid the ReNN learning to classify based on entity related features. We also tested for policy-generalizability of the classifier, by training it on three economic policies and evaluating its performance on the fourth economic policy, and similarly by training it on three technology policies and evaluating the performance on the fourth technology policy, and found it to perform reasonably well.

Finally, when we apply this system on the Indian mass media, we find that the media in general covers pro-policy statements more than anti-policy statements for economic policies, and technology deterministic statements more than skeptical views on technology policies. Our analysis can help build a media monitor, identify networks of support or conflict among the power elite, contrast statements made by them in the social media against their quotations highlighted in the mass media, and other interesting research questions that we plan to examine as part of the future work.

II. Related Work

Most work on ideological slant identification in the media has used sentiment analysis tools. These include generic dictionary-based tools like SentiStrength [7] and Vader [8], or specifically adapted for media settings by using partisan tokens [2, 9]. Although such approaches have been used to provide evidence of media bias by showing a correlation between the newspaper-slant and ideology of its readers [2], and showing how the bias affects voting patterns [9], such approaches are known to be coarse. Mullen et al. [10] show that such a traditional dictionary or phrase-based techniques are inadequate for political sentiment analysis. Yan et al. [11] on similar lines show that generalizing across different datasets or policies is also difficult since the concepts may be significantly distinct across different policies.

More recent approaches have used topic models to examine bias in news blogs, articles and speeches [12, 13]. Use of lexicons [14, 15] and accounting for POS (parts of speech) tags [16, 17] have also been applied on social media data. Other approaches use supervised machine learning models [18, 19, 20], Hidden Markov Model (HMM) based models [21], hierarchical topic modeling [22], and deep neural networks [23, 24] which are able to learn more complex nuances of language. Dong et al. [24] use an Adaptive Recursive Neural Network for entity-level Twitter Sentiment analysis where they propagate the sentiment from target-related sentiment words to the entity. However, these coarse-grained approaches may not always work well in stance detection of political *by-statements* where ideological bias is localised to a small portion in the sentence and hence structure of the sentence has to be taken into account. We build our approach based on the work by Iyyer et al. [23], to detect political ideology at the sentence level, using a Recursive Neural Network (ReNN) based model. We use the same model but take care in separating a neutral/non-neutral classifier from the pro/anti policy classifier, and similarly for technology-related policies. We also use a fine-tuned word2vec model to initialize the embedding layer and do not re-train this layer while training the classifier. Named-entities are removed from the corpus to make the classification output entity-independent.

III. Background

We analyse data on four Indian economic and technology policies, for which our goal is to identify statements covered in the mass media by prominent people.

Economic Policies: We choose four prominent policy topics - Aadhaar, Demonetisation, GST (Goods and Services Tax), and Farmers' protests, since they are recent, contentious, and with wide national ramifications [25]. *Aadhaar* is a national identity project that assigns a biometric-based unique identity number to every resident, and is meant to serve as a foundation for authentication and identification of citizens for access to government schemes, banking, health, and other services. The policy has been criticized due to a lack of attention paid to data security and citizens' privacy, and several issues to do with its faulty and hurried implementation [26]. *Demonetisation* was a disruptive move by the government, where overnight following an announcement by the prime minister, all INR 500 and 1000 currency notes were banned, with a view to curtail black money and counterfeit cash. It was widely criticized for the distress it caused to common

people who had to queue up for hours to exchange their old notes for new, and the liquidity disruption in the informal economy (almost all of agriculture, and small and medium scale enterprises) that primarily operates in cash [27]. The *Goods and Services Tax (GST)* is an indirect tax levied at each step of the value-chain, with a view towards formalizing the economy, and to replace an earlier regime of multiple taxes with a single tax code. GST has also faced a lot of criticism due to tedious compliance procedures, initial problems in its IT systems, and its complexity, which has been especially difficult for small businesses to manage [28]. The final topic of *Farmers' Protests* covers agricultural issues in general, specifically marked by a series of massive protests by farmers during 2017-18, in demand of better prices for crops, loan waivers, crop insurance, etc. [29].

Technology Policies: We similarly choose four prominent technology policies that emerged recently and are also of wide national importance: Cashless economy [30], Digital India [31], e-Governance [32], and Aadhaar [33]. The push towards a *cashless economy* envisions that most financial transactions will not be conducted through the exchange of physical currency, but rather through the transfer of digital information between the transacting parties. *Digital India* is a mission by the Indian government to connect rural areas with high-speed Internet networks. The *National e-Governance Plan (NeGP)* is an initiative of the government to make all government services available to the citizens of India via electronic media, instead of them having to fill up paper forms. Such a largescale embrace of digitization conveys the perspective of increased efficiency and inclusion of underserved populations in the information age, and has also been supported by Indian companies. However the top-down rollout of such schemes especially in the absence of digital literacy programmes for a population that might be coming online for the first time, has also led to issues like misinformation and rumours, technological failures and access problems that led to unfair denial of government benefits to many deserving citizens, phishing and financial fraud, etc.

IV. Data and system

We obtain the data for the policies described above, from a set of six English national newspapers in India: *The Times of India, Indian Express, The New Indian Express, Telegraph, Deccan Herald* and *Hindustan Times*. We have assembled a daily archive of all news stories by these newspapers, since 2011, as described in a paper by Sen et al. [34]. Daily crawlers download new articles and filter articles for the policies of interest by doing a simple keyword search. The keywords for each policy are listed in the supplementary material [33] and were obtained following an iterative query expansion process until saturation [34]. We obtained 22,302 articles on Demonetisation (Nov 2016 to Oct 2019); 13,908 articles on Aadhaar (2011 to 2019); 22,179 articles on GST (Jan 2011 to Oct 2019); 85,486 articles on Farmers' Protests (Nov 2016 to Oct 2019); and 23,432 articles (Jan 2014 to Oct 2019) related to technology policies. We then run these articles through a public Named Entity Extraction (NER) service called OpenCalais. This service identifies the places in an article where people, organizations, locations, etc. are mentioned. An entity resolution step then merges the same entities together [34]. We next describe the steps following this point.

***By-Statement* Extraction:** We first need to extract the statements made by the person-entities identified in the downloaded corpus of news articles. We do this by building a dependency parse tree using Stanford CoreNLP [35] to obtain the parts-of-speech (POS) tags and the dependency tree of each sentence in which an entity is referred. Some of these sentences may be quoting the person (*by class*), some sentences may mention something about the person (*about class*), and some may simply mention the person (*other class*). In this paper, we only examine the *by class* statements.

Sen et al. [34] have shown that rules based on POS tags and the dependency tree can be used to identify the class to which a statement belongs (by, about, or others). This is done by identifying important relationships between various subject (e.g. 'nsubj', 'csubj', etc.) and object tags (e.g. 'dobj', 'pobj'). For instance, consider this *by-statement*: *"A sincere attempt will be made to provide water to standing crops of our farmers", Siddaramaiah said.*, the entity (*Siddaramaiah*) is a subject, and the verb *said* is the main predicate. The dependency 'nsubj' (nominal subject) shows the relationship between the main predicate (*said*) and the subject (*Siddaramaiah*). If the entity appears as a subject, then it can either be a *by* or an *about* statement. Moreover, if the predicate connected by the 'nsubj' dependency of the entity has a certain type (like *said*, *claimed*, *told*, etc.), then it can be classified as a *by-statement*.

Dataset Annotation & Coding Schema: We take a subset of the *by statements* obtained in the previous step, and annotate them manually to build a training and test dataset for the subsequent steps to identify relevant statements, and obtain their ideological classification. We code these statements on economic policies to one of the five classes - *Non-Relevant, Pro, Anti, No stance,*

TABLE I: Stance identification using manual annotation (**P**-Pro, **A**-Anti, **N**-No-Stance, **B**-Balanced)

Policy	Relevant (P, A, N, B)	Non Relevant	Total
Aadhar	350 (169, 34, 146, 1)	42	392
Demonetisation	1063 (512, 259, 243, 49)	192	1255
GST	650 (292, 145, 167, 46)	31	681
Farmers	1792 (961, 505, 262, 64)	107	1899
Technology	812 (553, 115, 134, 10)	263	1075

Balanced stance using a coding schema that we describe further below. Similarly, we code the statements on technology policies to five classes as well.

We created a coding schema based on guidelines on qualitative content analysis [36, 37]. For each code, a set of examples and questions were listed for annotators to consider in making their decision. Context information was provided for each statement by giving the previous and next sentence as well. 15 annotators from among colleagues in our research group were first familiarized with our study and the different policies through two training workshops. They were then asked to code 100 statements per policy on an initial version of the coding schema, and point out ambiguities. Based on this feedback, the final coding schema was then prepared, and a new set of 100 statements were coded with 4 annotations per statement by different annotators. Inter-coder reliability (using Cohen's Kappa statistic [38]) between 0.75-0.79 was obtained for all the topics, which is considered quite reasonable. This final schema was then used by the annotators to code all the 5302 statements. The annotated dataset is described in Table I. Since we identified very few *balanced* statements, we ignore this class altogether and use the data for the rest of the four classes only. This dataset is likely to be useful to other researchers as well and can be downloaded from our GitHub link [39].

Relevance Filtering: While our choice of keywords is able to identify articles relevant to a policy, not all the *by statements* may be relevant. For example, a statement like *"I concede defeat and congratulate Ananth Kumar for his performance in this poll," Nilekani, the face of UPA's flagship Aadhaar programme, told PTI.* is not relevant to our study about whether it is pro/anti Aadhaar but gets listed because of the presence of the keyword *Aadhaar*. We, therefore, experimented with several methods for relevance classification, including TF-IDF [40], rules based on dependency parse trees, and machine learning based methods trained on the annotated dataset described above. The results are given in more detail in the supplementary material [33]. Machine learning based methods gave us the best performance, and we finally chose a random forests classifier that gave an F1-Score of 0.95, 0.92, 0.98 and 0.92 for economic policies of Aadhaar, Demonetisation, GST and Farmers' Protests respectively, and 0.80 for Technology policies.

V. Ideology Detection and Classification

In this section, we describe our framework to detect the political ideology of a particular *by-statement*. The relevant statements obtained in the previous step are first passed to an Ideology Detection classifier which determines whether the statement holds any stance concerning the policy or not, and then to an Ideological Stance classifier (which classifies whether non-neutral statements are in support of the policy or against it). There are two versions of both these classifiers, for economic policies and technology policies, respectively. All four classifiers use the same architecture and training procedure, which we now describe.

Model: We use a Recursive Neural Network (ReNN) based architecture inspired by Iyyer et al. [23], as our classification model. Note that the ReNN approach we have adopted is different from a typical Recurrent Neural Network (RNN) as ReNN is more amenable to learning from smaller datasets since it imposes a hierarchical decomposition of sentences into smaller phrases. A generic RNN would need more data to arrive at a similar model. ReNN works on the assumption that the meaning of each phrase would be a combination of the meaning of the words that form it, and the syntax that combines these words together. To break a sentence into phrases, we use the Stanford CoreNLP parser to obtain the parse tree and feed it to the model as an input. Although different phrases in a sentence may also have different ideological stances that combine to reflect the overall ideology portrayed by a sentence, a limitation of our work so far is that we do not annotate each phrase separately, rather we assume that all the constituent phrases would have the same stance as that of the sentence overall. Despite this simplification to not handle the complexity in sentences, we are able to achieve reasonable performance, and also generalizability across policies (section VI).

To represent the phrases of a parsed sentence as vectors, embeddings of words forming a particular phrase are combined to build a *phrase vector* (figure 2) that

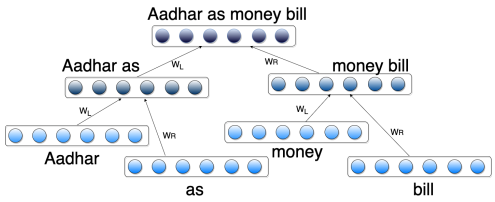


Fig. 2: Example of how word representations are combined to form phrase vectors of the same dimensions

has the same dimensions as the word embeddings. If two words w_a and w_b combine to form a phrase p, then the vector representation of the phrase x_p is given by: $x_p = f(W_L.x_a + W_R.x_b + b_1)$ where x_a and x_b are the word embeddings of w_a and w_b, derived from an embedding matrix W_e of dimension $d \times V$, V being the size of the vocabulary. f is a non-linear activation function, W_L and W_R are the left and right composition matrices, and b_1 is a bias term. The ideology of each phrase is then calculated as: $\hat{y}_p = \text{softmax}(W_{cat}.x_p + b_2)$ where W_{cat}, b_2 are parameters. We use a cross-entropy loss function for training. We also use $L2$ regularisation to avoid overfitting, given the small size of our dataset. The parameters of the model are optimized using Stochastic Gradient Descent with momentum. To evaluate the performance of the overall model, we use the macro-averaged F-score $F_{macro} = \frac{F_{Neutral} + F_{Non-Neutral}}{2}$ at Step-1 (stance detection) and $F_{macro} = \frac{F_{Pro} + F_{Anti}}{2}$ at Step-2 (ideology classification).

Word Representation: We use the word2vec embedding matrix [41] pre-trained on the Google News corpus [42] to initialize the embedding layer of the model. In section VI, we show that pre-trained word2vec embeddings do not perform as well as when the embeddings have been fine-tuned using the news articles in our dataset. This is because many words such as "digital" or "smart" have a specific meaning in our domain, usually referring to *"Digital India"* or *"Smart Cities"*, hence fine-tuned embeddings are able to learn these patterns. We build two sets of fine-tuned word2vec embeddings, for economic policies and technology policies, respectively.

When training the ideology models based on these fine-tuned word2vec embeddings, we again have a choice of whether to re-train the embedding layer along with the whole model or to freeze the embedding layer while we train the rest of the model. We choose to do the latter due to the small size of our annotated dataset. This choice is also supported by the results in section VI, which show that the frozen embedding performs better.

Making classification entity-independent: We also discovered that the presence of names of entities in the sentences could lead to misclassifications. For example, the statement *"PM Modi had announced the note ban on November 8 last year, and the decision destroyed the country's economy."* is clearly an *anti-policy* statement but it was misclassified as *pro-policy*, due to occurrence of the dominantly pro-policy entity of "prime minister Modi" in the sentence. We, therefore, blind out the entities in our training dataset by replacing all entities with a common token, and as shown in section VI, we find that this results in a better performance. We refer to this final model as **ID-ReNN** (Ideology Detection - Recursive Neural Network). Figure 3 further qualitatively validates our decision of blinding the named entities before fine-tuning. It shows a t-SNE plot of various named entities in the word-embedding space, without blinding. The blue dots represent entities with a dominant *pro* stance while the red dots represent those with a dominant *anti* stance. We can see that *pro* entities mostly belong to the ruling coalition (belonging to the BJP - Bhartiya Janata Party) whereas those against are from the opposition. Moreover, BJP leaders are found to be in close proximity of each other, implying that they get mentioned in similar sentences; same is the case with opposition leaders, who form separate clusters. We also found that words used to support a policy, like "anti-corruption" and "cashless", have a greater association (measured using cosine-similarity) with BJP entities as compared to opposition entities. These types of dominant entity associations ultimately affect the output of the classifier negatively, and it is better to blind them out so that the classifier just learns to classify based on the sentence structure.

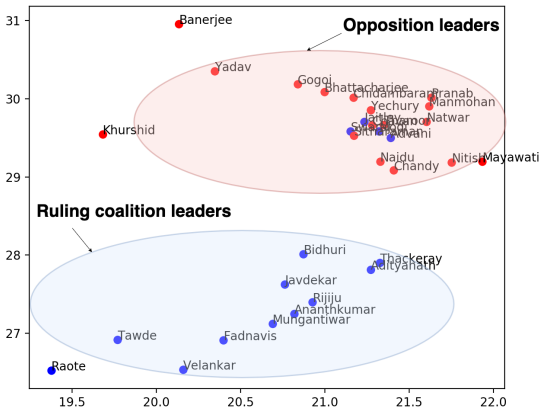


Fig. 3: t-SNE visualisations of various entities

VI. Results

Table II gives an overview of our results, in comparison with other models. All the experiments are done by undersampling the majority class or oversampling the minority class, to account for class imbalance, as indicated in the table. The low accuracy of SentiStrength [7] as a baseline is expected as it is not robust to various sentence structures, and hence fails to capture important semantics like sarcasm, negation, indirect words, etc. at the phrase level. We also experiment with several state-of-the-art neural network architectures as a baseline, out of which we obtain our best results with BERT (Bidirectional Encoder Representations from Transformers)

TABLE II: Performance comparison of our model (**ID-ReNN**) (**U**: Undersampling, **O**: Oversampling, **Acc**: Accuracy)

Models	Step-1 (Stance Detection) Neutral vs Non-Neutral				Step-2 (Ideology Classification) Pro vs Anti			
	Economic		Technology		Economic		Technology	
	U	O	U	O	U	O	U	O
	Acc% (F1)	Acc% (F1)	Acc% (F1)	Acc% (F1)	Acc% (F1)	Acc% (F1)	Acc% (F1)	Acc% (F1)
SentiStrength	43.5 (0.42)	43.2 (0.42)	41.2 (0.38)	44.2 (0.38)	60.1 (0.58)	58.7 (0.59)	64.1 (0.63)	64.3 (0.67)
Iyyer et al	61.5 (0.63)	64.1 (0.67)	73.0 (0.73)	74.6 (0.75)	63.1 (0.66)	65.8 (0.70)	72.4 (0.73)	76.2 (0.76)
Fine-tuned BERT	66.1 (0.61)	77.4 (0.66)	67.8 (0.60)	81.0 (0.60)	76.6 (0.75)	78.7 (0.76)	80.8 (0.7)	**91.9 (0.85)**
ID-ReNN	**78.1 (0.75)**	**78.5 (0.77)**	**84.6 (0.80)**	**86.7 (0.90)**	**78.9 (0.79)**	**80.1 (0.81)**	**85.5 (0.82)**	90.7 (0.93)

[43] fine-tuned on our corpus. Other neural network baselines are given in the supplementary material [33]. We also experimented with machine learning algorithms including Linear SVMs, Random Forests (RFs), k-Nearest Neighbors (kNN), and Naive Bayes (NB), but did not get as good results as the neural network models. Our method works better than fine-tuned state-of-the-art BERT (Base-Uncased) for the specific task of identifying the ideology of *by-statements*.

Our model (referred to as **ID-ReNN**) is evaluated under different settings: (a) ID-ReNN-G: Use of generic word2vec embeddings based on Google News Corpus, (b) ID-ReNN-FT: word2vec embeddings fine-tuned on a collection of news articles from our corpus on economic and technology policies, and (c) ID-ReNN-FT-N: Fine-tuned embeddings obtained using our corpus and replacing the *Person* and *Organization* type entities with a common blinding token. In addition to experimenting with the different types of embeddings, we also experiment with training the model by allowing the input layer weights to re-train (represented by suffix (T) to the model name), or by keeping them frozen (represented by suffix (F) to the model name). The performance of our model with these different variations is shown in table III. The best performance was obtained by using ID-ReNN-FT-N(F) as explained earlier.

TABLE III: Model performance in different settings

Model	Economic		Technology	
	Accuracy	F1	Accuracy	F1
ID-ReNN-G(F)	69.8%	0.76	80.2%	0.81
ID-ReNN-G(T)	65.8%	0.70	76.2%	0.76
ID-ReNN-FT(F)	71.6%	0.78	82.8%	0.71
ID-ReNN-FT(T)	67.7%	0.71	78.4%	0.69
ID-ReNN-FT-N(F)	**80.1%**	**0.81**	**90.7%**	**0.93**
ID-ReNN-FT-N(T)	73.7%	0.76	85.6%	0.88

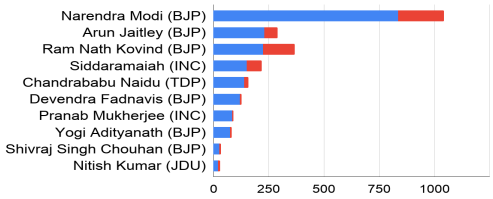
To investigate the generalizability of the ideology classifier for economic policies, we trained our model on three policies and tested on the fourth policy. We achieved a test performance (on the unseen policy) of 74.8% (F1-0.78) for Aadhaar, 70.7% (F1-0.72) for Demonetisation, 76.2% (F1-0.82) for GST and 65.8% (F1-0.72) for Farmers' Protests, after training on the other three policies in each case. Our model seems to perform reasonably well, other than for Farmers' Protests, which may be because this particular policy topic is indeed significantly different from the other topics. For the technology classifier, we achieved a test performance of 89.0% (F1-0.81) for Aadhaar, 82.3% (F1-0.75) for Cashless Payments, 84.6% (F1-0.77) for Digital India, and 88.1% (F1-0.70) for E-Governance.

VII. Analyzing Ideological Biases

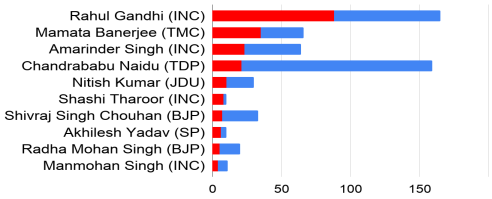
The proposed system can have various applications related to ideological bias analysis. We answer two research questions in this direction: (a) Which people are the most supportive or critical of the economic and technology policies on mass media, and (b) Do newspapers display a dominant ideological slant based on their coverage of statements by important people. Our work is related to that by Sen et al. [34], which analyzes biases in the Indian media based on the coverage given to different aspects of a policy. Here, we examine bias based on the ideological slant of statements covered by mass media. This builds upon works such as that by Budak et al. [18] who use manually annotated articles in the US media, and show that a common strategy for newspapers to introduce negative bias against somebody is by giving greater coverage to critical or negative statements made by the opposition. We study two sets of policies w.r.t. these questions, namely economic and technology policies.

Economic Policies:

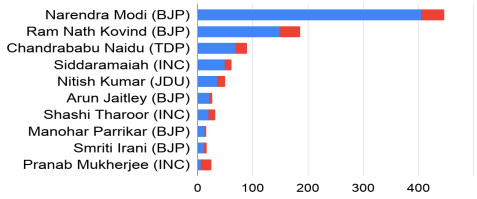
Ideological Position of Entities: To estimate the ideological position of an entity, we obtain the counts of *Pro* and *Anti* statements made by that entity. The ideological position of the top 10 entities is shown in figure 4a and 4b. We find prominent politicians from the ruling Bharatiya Janta Party (BJP) (such as the prime minister *Narendra Modi* and the late finance minister *Arun Jaitley*), and the opposition parties (such as *Rahul Gandhi* from the Indian National Congress and *Mamata Banerjee* from the Trinamool Congress) among the top entities. As expected, the ruling politicians mostly make pro-policy statements, whereas the opposition makes anti-



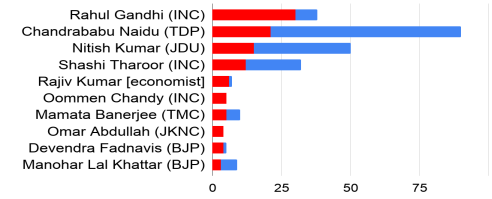
(a) Pro Economic Entities



(b) Anti Economic Entities



(c) Tech Deterministic Entities



(d) Tech Skeptic Entities

Fig. 4: Top 10 Entities (in terms of Pro/Anti statements) for both economic and technology policies – political affiliation within braces (.); Blue and red labels denote the number of pro and anti statements respectively

policy statements. We use chi-square tests to check this relationship between the party affiliation of entities and their stance and find statistically significant dependence of BJP entities with pro and Non-BJP entities with anti.

Ideological Slant of Mass Media: Table IV shows the pro/anti distribution of statements covered by different mass media sources. We see that across all sources, the proportion of pro-policy statements significantly exceeds that of anti-policy statements. On examining the sentiment slant of the corresponding articles (table V) using SentiStrength[1], we find that all of the media houses report quite negatively on economic policies (sentiment slant below -1), by heavily quoting politicians who support them in their articles (as evident from figure 4), and countering their claims simultaneously. We plan to study this strategy of criticism in detail in the future.

Technology Policies:

Ideological Position of Entities: For technology policies, we again find that the ruling politicians are quoted for their favourable views on the policies, whereas the opposition is quoted for expressing skepticism. The findings are also in agreement with earlier studies [44, 45] which show a technology deterministic ideology prevalent among policymakers.

Ideological Slant of Mass Media: We find that the mass media gives substantial coverage to statements in favour of the technology policies (table IV). An article-level sentiment analysis reveals a dominant tech-deterministic coverage in most media houses (as shown in table V) as corroborated by [34]. It seems to suggest that not only do most media houses cover technology policies positively,

[1]Based on a manual analysis of 200 articles (class balanced) for each economic policy by three annotators, the accuracy of SentiStrength was found to be in the range of [75%,80%]

TABLE IV: Pro/Anti statement classification distribution in media. T-tests prove the hypothesis that the newspaper sources give significantly more coverage to pro-statements by entities. (**Det:** Deterministic)

Newspaper Source	Economic (%)		Technology (%)	
	Pro	Anti	Det.	Skeptic
Deccan Herald	73.6	26.4	85.1	14.9
Hindustan Times	70.6	29.4	84.5	15.5
Indian Express	74.5	25.5	84.6	15.4
New Indian Express	68.1	31.9	80.0	20.0
The Times of India	68.2	31.8	86.3	13.7
Telegraph	70.0	30.0	81.9	18.1

TABLE V: Article level sentiment analysis done using SentiStrength for 1000 randomly sampled articles per policy for each newspaper. Columns denote the percentage of highly negative articles (those with sentiment values below -1 in the range of -5 to +5.)

Newspaper Source	Economic	Technology
	% Anti Articles	% Skeptic Articles
Deccan Herald	56.67	34.00
Hindustan Times	80.08	59.21
Indian Express	80.26	57.81
New Indian Express	64.63	39.85
The Times of India	54.87	34.64
Telegraph	64.81	43.75

but they also provide more attention to the politicians having a tech-deterministic standpoint, thereby supporting the technology determinism that is already prevalent in the government, and among the business-persons. These findings are also corroborated by earlier works [34, 46], where Thrall et al. [46] show preferential coverage of interest groups having more resources.

VIII. Conclusion

We demonstrated a framework to study ideological biases based on the coverage given by the newspapers

to statements made by prominent people. We use a Recursive Neural Network based model to classify the statements into three classes of pro-policy, anti-policy, and neutral ideologies, for economic and technology policies. Our findings indicate that the Indian news-sources generally cover pro-policy statements much more than those criticizing them, and takes a tech-deterministic standpoint on technology policies. The end-to-end system developed by us can serve to answer interesting research questions in future to compare social media self-expression by prominent people with the coverage provided to these statements selectively by the mass media.

REFERENCES

[1] M. Mccombs, "The agenda-setting role of the mass media in the shaping of public opinion," 01 2011.

[2] M. Gentzkow and J. Shapiro, "What drives media slant? evidence from u.s. daily newspapers," *Econometrica*, 2010.

[3] J. Milyo and T. Groseclose, "A measure of media bias," *The Quarterly Journal of Economics*, vol. 120, pp. 1191–1237, 2005.

[4] C. W. Mills, *The Power Elite*. Oxford University Press, 1956.

[5] R. Williams and D. Edge, "The social shaping of technology," *Research Policy*, vol. 25, pp. 865–899, 1996.

[6] X. Fang and J. Zhan, "Sentiment analysis using product review data," *Journal of Big Data*, vol. 2, no. 1, p. 5, 2015.

[7] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," 2010.

[8] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," 2015.

[9] S. Gerrish and D. Blei, "Predicting legislative roll calls from text." 2011, pp. 489–496.

[10] T. Mullen and R. Malouf, "A preliminary investigation into sentiment analysis of informal political discourse." 2006.

[11] H. Yan, A. Lavoie, and S. Das, "The perils of classifying political orientation from text," in *LINKDEM@IJCAI*, 2017.

[12] A. Ahmed and E. Xing, "Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010, pp. 1140–1150.

[13] W.-H. Lin, E. Xing, and A. Hauptmann, "A joint topic and perspective model for ideological discourse," 2008, pp. 17–32.

[14] P. Ray and A. Chakrabarti, "Twitter sentiment analysis for product review using lexicon method," in *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, 2017, pp. 211–216.

[15] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," vol. 11, 2010.

[16] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," vol. 2, 2010, pp. 36–44.

[17] I. Weber, V. Garimella, and A. T. Hadgu, "Political hashtag trends," vol. 7814, 2013, pp. 857–860.

[18] C. Budak, S. Goel, and J. Rao, "Fair and balanced? quantifying media bias through crowdsourced content analysis," *Public Opinion Quarterly*, vol. 80, pp. 250–271, 2016.

[19] D. Preotiuc-Pietro, Y. Liu, D. Hopkins, and L. Ungar, "Beyond binary labels: Political ideology prediction of twitter users," 2017.

[20] M. Lai, A. Cignarella, and D. Hernandez Farias, "itacos at ibereval2017: Detecting stance in catalan and spanish tweets," 09 2017.

[21] Y. Sim, B. Acree, J. Gross, and N. Smith, "Measuring ideological proportions in political speeches," *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 91–101, 2013.

[22] V.-A. Nguyen, J. Boyd-Graber, and P. Resnik, "Lexical and hierarchical topic regression," *Advances in Neural Information Processing Systems*, 2013.

[23] M. Iyyer, P. Enns, J. Boyd-Graber, and P. Resnik, "Political ideology detection using recursive neural networks," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 1113–1122.

[24] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent twitter sentiment classification," 2014.

[25] "Gautam chikermane. 2018. nine economic policies that define modi@4." https://www.orfonline.org/expert-speak/nine-economic-policies-that-define-modi-4/.

[26] J. Drèze, N. Khalid, R. Khera, and A. Somanchi, "Aadhaar and food security in jharkhand: Pain without gain?" *Economic and Political Weekly*, vol. 52, pp. 50–60, 2017.

[27] N. K. Krishnan, A. Johri, R. Chandrasekaran, and J. Pal, "Cashing out: digital payments and resilience post-demonetization," 2019.

[28] A. Das-Gupta, "Some problems with the indian goods and services tax," *SSRN Electronic Journal*, 2018.

[29] "Epw engage. 2018. why are our farmers angry?" https://www.epw.in/engage/article/farmer-protests-delhi.

[30] Wikipedia contributors, "Cashless society — Wikipedia, the free encyclopedia," 2020, https://bit.ly/34W6uaq.

[31] ——, "Digital india — Wikipedia, the free encyclopedia," 2020, https://bit.ly/3eGDFD5.

[32] ——, "National e-governance plan — Wikipedia, the free encyclopedia," 2020, https://bit.ly/2XQN1WV.

[33] "Supplementary material: Ideology detection in the indian mass media," https://tinyurl.com/y4zc5k8d.

[34] A. Sen, P. Chhillar, P. Aggarwal, S. Verma, D. Ghatak, P. Kumari, M. Agandh, A. Guru, and A. Seth, "An attempt at using mass media data to analyze the political economy around some key ictd policies in india," 2019, pp. 1–11.

[35] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," http://www.aclweb.org/anthology/P/P14/P14-5010.

[36] G. A. Amedeka, "Newspaper coverage of the 2010 district assembly election in ghana: A content analysis of daily graphic and daily guide," Ph.D. dissertation, University of Ghana, 2015.

[37] K. Smith, M. Wakefield, C. Siebel, M. Szczypka, S. Slater, and S. Emery, "Coding the news: The development of a methodological framework for coding and analyzing newspaper coverage of tobacco issues," 2002.

[38] M. McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, vol. 22, pp. 276–82, 2012.

[39] Ankur Sharma, "Annotated dataset," 2020, https://bit.ly/315uo1q.

[40] H. C. Wu, R. Luk, K.-F. Wong, and K.-L. Kwok, "Interpreting tf-idf term weights as making relevance decisions," *ACM Trans. Inf. Syst.*, vol. 26, 2008.

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[42] "Google archive (word2vec)," 2013, https://bit.ly/3fkU5QC.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[44] J. Pal, "The technological self in india: From tech-savvy farmers to a selfie-tweeting prime minister," 2017, pp. 1–13.

[45] J. Pal, P. Chandra, V. Kameswaran, A. Parameshwar, S. Joshi, and A. Johri, "Digital payment and its discontents: Street shops and the indian government's push for cashless transactions," 2018.

[46] T. Thrall, "The myth of the outside strategy: Mass media news coverage of interest groups," *Political Communication*, vol. 23, pp. 407–420, 10 2006.