EasyChair Preprint

№ 4591

Ideology Detection in the Indian Mass Media

Navreet Kaur and Ankur Sharma

November 18, 2020

Ideology Detection in the Indian Mass Media

Thesis submitted by

Navreet Kaur **Ankur Sharma**

2015TT10917 **2015CS50278**

under the guidance of

Prof. Aaditeshwar Seth

in partial fulfilment of the requirements
for the award of the degree of

Bachelor and Master of Technology

July 2020

Department Of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY DELHI

THESIS CERTIFICATE

This is to certify that the thesis titled **Ideology Detection in the Indian Mass Media**, submitted by **Ankur Sharma** and **Navreet Kaur**, to the Indian Institute of Technology, Delhi, for the award of the degree of **Bachelor and Master of Technology**, is a bonafide record of the research work done by them under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Aaditeshwar Seth
Professor
Department of Computer Science
Indian Institute of Technology, Delhi

ACKNOWLEDGEMENTS

Foremost, we would like to express our sincere gratitude to our advisor Prof. Aaditeshwar Seth for the continuous support of our Master's study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped us in all the time of research and writing of this thesis. We could not have imagined having a better advisor and mentor for our Master's study. He consistently allowed us to frame our work but steered us in the right direction whenever he thought we needed it.

We would also like to thank Anirban Sen for his immense support and guidance during the past year. He brought structure to both our research and thinking process, allowing us to deep dive and grow into this field of study. We will cherish our late-night sprints during our paper submissions and our stimulating discussions regarding minute details of our algorithm or codebase. Anirban consistently brought out the best in us while giving us our own space to work and think, leading us to believe we have become better researchers by the end of it.

No research is possible without infrastructure and necessary resources. For this, we extend thanks to HPC facility and other resources provided by ACT4D and IIT Delhi. Big thanks to our colleagues, the members of the ACT4D research group at IIT Delhi, especially for giving numerous hours of their time to help us annotate the datasets.

Navreet: It's my fortune to gratefully acknowledge the support of my friends Ankur Sharma, Aman Singh, Aditya Sahdev, Madhav Ranka and Sanyam Gupta for their support, care and valuable discussions throughout the research. They were always beside me during the happy and hard moments to push me and motivate me.

Ankur: I would also like to acknowledge my friends and labmates at IIT Delhi: Dishant Singla, Sudeep Agarwal, Lovish Madaan and Chakshu Goyal for being my right-hand support throughout all my course work and assignments, and I am gratefully indebted to their valuable comments on this thesis. A heartful thanks to Yashaswini Gupta at Ambedkar University (Delhi) for lending her brilliant proofreading skills to my publications. I am much indebted to my friends: Navreet Kaur, Sanyam Gupta, Saksham Gupta, Mallika Singla, Mehak Preet and Arshiya Bhutani, whose dedication, love and support have helped me a lot especially during my hard time with the extraction work. I owe them for being unselfishly helping me throughout the research. They were always beside me during the happy and hard moments to push me and motivate me.

Last but not least, we would like to thank our family: our parents for supporting us

spiritually throughout our lives. We salute them for all for the selfless love, care, pain and sacrifice they did to shape our lives. Although they hardly understood what we worked on, they were willing to support the decisions we made. We are grateful to them for cultivating our curiosity and pushing us to be good human beings along our journeys.

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.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Motivation

Ideology is a set of beliefs that is carried by a person or a group. It relates to the preferred behavioral options undertaken during social decision making. Several studies [12, 25, 28] have proven that mass media might be biased towards or against a particular government, political party, or ideology.

Why do we need to study media bias?

Assessing media bias is thus an essential need since the mass media is known to significantly shape public opinion [22] and the social context in which the policies are developed. In this paper, we study the ideological bias carried by the Indian mass media, in terms of the statements that it covers on important economic and technological policies. These statements are often made by influential people (or entities) who might be in favor of or against a policy. In other words, these statements help us understand the ideological positions of the entities on policy issues. The preferential coverage provided to these entities and their statements, in turn, helps us understand the ideological bias of the mass media.

Why do existing approaches fail?

For political statements, such ideological positions may be localized to a particular part of the sentence. Existing approaches on sentiment analysis prove to be insufficient for this task of ideology detection, since they often fail to pick up complex linguistic features that explain ideological positions, like sentence structures, negations, and contextual information. It becomes increasingly difficult if the sentence contains sarcasm and the use of contrasting sentiment words. For example, SentiStrength, a popular tool for sentiment analysis, classifies the statement, *Just because it is possible to hack a network does not mean that technology must not be deployed.* as a negative sentiment statement even though it is a pro-technology statement, which can be because of the use of multiple negations here. Similarly the statement *Earlier farmers used to get insurance of Rs 50,000 on death or permanent disability under Raj Sahakar Personal Insurance Scheme, which now has been increased to Rs 10 lakh.* requires information of the domain of discussion to predict that increasing the insurance amount for farmers is a positive outcome. Please refer to table 1.1 which shows some examples in which the Senti Strength apporach actually fails.

The failure of sentiment analysis techniques in the policy domain may stem from the fact that both sides of the ideological stance – positive and negative – generally use the same

Table 1.1: Failure of SentiStrength

Entity Statement	**Actual Stance**	**Senti Strength Score**	**Explanation for SentiStrength Failure**
In order to help children suffering from an e-learning tool has also been developed by the National Institute of Mentally Handicapped and all the scholarships will be under one	Pro	-3	Mentions word 'Autism'
Bangalore has enormous energy and ideas to solve problems.	Pro	-1	Mentions word 'problems'
At UIDAI, we are very strict on privacy issues.	Pro	-2	Mentions word 'strict'
Earlier farmers used to get insurance of Rs 50,000 on death or permanent disability under Raj Sahakar Personal Insurance Scheme, which now has been increased to Rs 10 lakh.	Pro	-2	Mentions words, 'death' and 'disability'
Prime Minister Narendra Modi 's grand MSP increase for farmers is like applying a band-aid to a massive hemorrhage.	Anti	1	Sarcasm

set of words while making a statement about the policy. Hence, ideology detection of such statements requires us to understand the context of the sentence rather than considering the sentence merely as a bag-of-words.

1.2 Overview

In this work, we use natural language processing and deep learning to analyze Indian mass media articles on major economic and technological policy events. We study important economic policies – Aadhaar, Demonetisation, GST, and Farmers' Protests to understand how various newspapers preferentially cover specific aspects and certain entities involved in these policies. We also study technology policies as a separate group in the same way.

Our ideology detection framework has been shown in figure 1.1. It consists of three main components:

1. **Data Extractor**: This module extracts statements made by different entities from articles on specific policies. Domain-specific keywords are used to extract news articles related to certain policies, and a rule-based *by-statement* extractor is used to extract

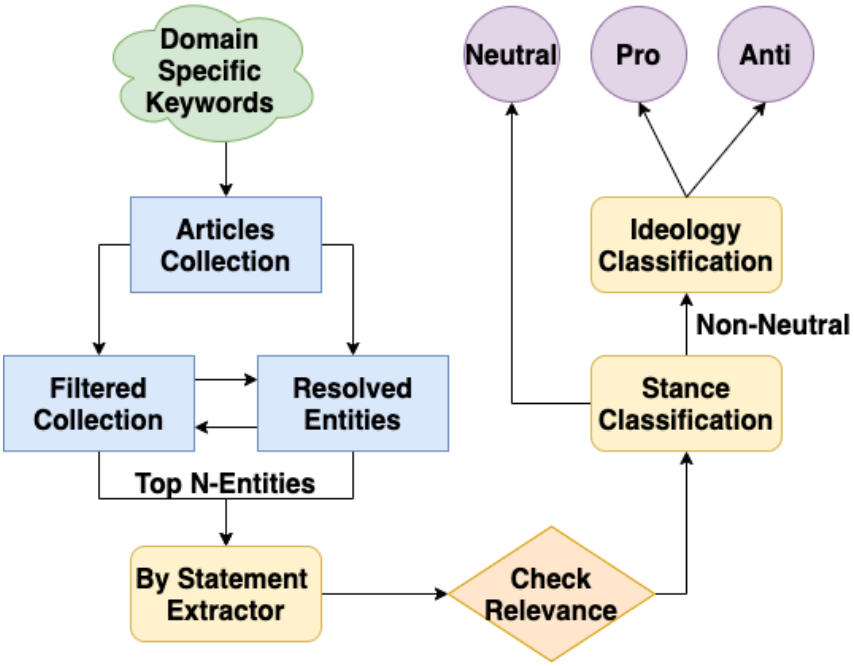


Figure 1.1: Our proposed Ideology Detection Framework

statements made by different entities from these articles.

2. **Relevance Filter**: This module filters out the non-relevant statements extracted in the previous step.

3. **Classifier**: This module consists of a two-step classifier that first checks if a statement is neutral, and then checks if the non-neutral statement is in favor of (pro) or against (anti) the policy.

We develop a two-step classifier using a recursive neural network at each step. Our classification approach is entity-independent, i.e., it does not depend on the affiliation or ideology of the entity (like a politician affiliated to a certain political party) that makes the statement. For this purpose, we use a method of preprocessing the data to obtain fine-tuned word vectors whose meanings are not associated with the ideology or affiliation of an entity. Our aim is to build an umbrella classifier that can be applied to different datasets or policies. Hence, we adopt a method of training procedure that ensures generalizability on new policies. We do this by freezing the embedding layer of the model while training, and also ensure that the dataset used for training the model has enough domain information for the classifier to learn various common axes of economic or technology policy debate and evaluate any other economic or technology policy on these same axes. The details of the model, the fine-tuning approach, and the training process are described further in chapter 6.

1.3 Contributions

With this goal in mind, we propose a method to automatically extract *by-statements* relevant to a policy from multiple news-sources and release a dataset of statements made by various entities, about the five mentioned policies, published in six national newspapers. We also propose a comprehensive framework to examine and analyze the political ideologies favored by media automatically. This framework consists of three key steps - Article Extraction, Relevance Check and Ideology Classification.

In other words, our primary contributions are:

1. Two annotated datasets containing 3855 and 812 statements related to economic and technological policies, respectively (refer to table 4.4 for details).
2. Two fine-grained stance detection models to study if a statement is in favor of or against an economic or technological policy, and their application on the Indian mass media data to understand the underlying ideological bias accurately. We also show that our stance detection models are generic enough to be applied to new policy data as well.

We use this framework to study the ideological biases in the Indian mass media in terms of the policy-related statements that they cover. Our findings show that the Indian mass media typically covers pro-policy statements much more than anti-policy statements, and covers pro-technology viewpoints much more than the other side of the discourse, indicative of an ideological bias existing in mass media. We also demonstrate how our framework is generic enough to be applied to any other domain of ideology classification.

1.4 Thesis Outline

The remainder of our thesis is organised as follows. Chapter 2 contains the Related Work, followed by the chapter 3 on the domain/background description of policies. Chapter 4 contains the system design, architecture and algorithms for data collection and *by-statement* extraction. We also explain our media analysis database and our website in this Chapter. We enlist our target problem in detail in chapter 5. Our methodology for stance classification is described in 6. Chapter 7 and 8 contains our results and analysis on the economic and technology policies. Chapter 9 contains the conclusion and the future scope of work in this project.

Chapter 2

RELATED WORK

There has been extensive work on political-ideology detection through natural language processing. Various computational frameworks have been developed to automatically detect political ideology of news articles, social media posts, and parliamentary speeches [14]. In this section, we highlight studies related to our work.

2.1 Bias in Mass Media

Mass media has significant effect on policy formulation [22] and is known to shape public opinion by intentionally having bias in their writing, coverage and distribution of news, which is why they are often referred to as *gatekeepers* [37]. Scheufele et al. discuss the concepts of agenda setting, framing, and priming in mass media [32], which together play a significant role in influencing public opinion on socio-political issues. [32] observe how content targeted at selective groups of audience to maximize profitability by catering to advertiser interests affects public policies, because it amplifies the outreach of politicians to these audience groups. Oliver and Myers [29] write that the claim of the media being an objective and neutral communicator of events has been rejected by scholars of the media for quite some time now. In the study [6], Bartels finds a significant influence of mass media exposure in opinion shift of the public in the 1980 US presidential elections.

Our work is related and aims to understand the ideological bias existing in Indian mass media, in terms of the coverage it provides to the statements given by influential people on key economic and technological policies.

2.2 Detecting Ideology

Sentiment analysis techniques have been used in some works to identify the ideology of a particular statement. The traditional methods include use of partisan tokens [12] and bag-of-words [13] for ideology detection. In one of the earliest works in this field, Gentzkow and Shapiro [12] developed a "slant-index" that quantifies media slant, by analyzing key phrases in the news content specific to political ideologies. Apart from a domain-specific phrase or word detection, researchers have also leveraged various sentiment analysis tools like Sentistrength, Vader, Alchemy, etc. to analyze the sentiment exhibited in text. While

these works give strong evidence of media being biased, by showing a correlation between newspaper-slant and ideology of its potential readers [12], and showing how media bias affects voting pattern [13], such approaches are too coarse to understand the political ideology resulting from the use of complex evidence and words in a particular context. These tools are based on the presence or absence of certain words which are clearly indicative of a certain sentiment, and ignore the sentence structures, negations, and contextual information, which also play a key role in determining the overall stance of the statement. For example, the word "good" contributes towards a positive sentiment, and the word "bad" contributes towards a negative sentiment. Quite often, such tools ignore the sentence structures, negations, and contextual information, which also play a key role in determining the overall stance of the statement. Most of the conventional techniques essentially focus on bag-of-words models which ignore the syntax. Mullen et al. [26] show that such traditional text classification techniques are inadequate for the task of political sentiment analysis. This can be attributed to the arbitrariness in the statements belonging to a particular class of policies. Yan et al. [47], on similar lines, show that generalizing across different datasets or policies, and making an umbrella classifier is extremely difficult as concepts are significantly distinct across policies.

More sophisticated approaches towards political sentiment analysis include Hidden Markov Model (HMM) based models [38] and hierarchical topic modeling [27]. Sim et al. [38] propose an HMM-based model, which uses a fixed lexicon of bigrams to infer the ideology used by political candidates in their campaigns. Inspired by a two-level political science theory, which unifies agenda setting and ideological framing, Supervised Hierarchical Latent Dirichlet Allocation (SHLDA) [27] is seen to improve prediction of political affiliation and sentiment. More recent works include the use of machine learning and deep neural networks for natural language processing, which have proved to be effective in incorporating the complicated nuances of language. They therefore predict the correct sentiment accurately. For example, Budak et al. [9] use crowd-sourcing and machine learning techniques to understand whether or not the US media reports in a non-partisan manner. Iyyer et al. [14] have used recursive neural networks (RNNs) to detect political ideology at the sentence level. Inspired from their work, we have built a Recursive Neural Network based model for ideology detection in economic and technological policies. The difference of our work from the aforementioned studies is that we develop a two-step classifier with a training procedure that ensures generalizability on new policies, and a method of fine-tuning to initialize the embedding layer, which makes classification entity-independent.

Chapter 3

BACKGROUND

In our study, we analyze mass media bias corresponding to four economic policies, and a group of technological policies hereinafter referred to as *technology policies*. We build two classifiers to identify the ideology of a statement - *Pro* versus *Anti* for economic policies, and *Determinism* versus *Skepticism* for technology policies. The detailed definitions of these terms can be found in section 5.

3.1 Economic Policies

The economic policies that we study in this work are Aadhaar, Demonetisation, GST, and Farmers' protests since these are recent, contentious, and of national importance [3].

3.1.1 Aadhaar

Under the Aadhaar policy, the government took the initiative to assign every Indian resident a biometric-based unique identification number [11]. The data is collected by the Unique Identification Authority of India (UIDAI), a statutory authority under the jurisdiction of the Ministry of Electronics and Information Technology. The policy has been criticized by some politicians, social activists, policy experts, and economists 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, and illicit use of the platform by different agencies.

3.1.2 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 [18]. 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, and cash exhaustion in ATMs, among other reasons. The suddenness of the policy move also led to suffering on part of the farmers and daily wage earners due to shortage of cash among people, and understaffed banks unable to dispense cash.

3.1.3 Goods and Services Tax (GST)

GST is an indirect tax levied in India on the sale of goods and services [10]. 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 on its complexity and problems in implementation, which have impacted the overall growth of the economy.

3.1.4 Farmers' Protests

This policy issue covers series of protests [2] 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, forest rights, and other issues related to agriculture in general. The issue is highly active politically with significant involvement of different politicians and political parties.

We select these economic policies for our analysis as they are some of the key policies implemented or taken forward by the current government, as reported in multiple media outlets and forums [3].

3.2 Technology Policies

Technology policies refer to a group of policies that include various technical interventions aimed at solving problems of the people. It includes several key policy issues like Cashless Economy [42], Digital India [43], E-Governance [44], and Aadhaar [11].

3.2.1 Aadhaar

Aadhaar [11] is a 12-digit unique identity number that can be obtained by residents of India, based on their biometric and demographic data. Aadhaar has been positioned as a tool that can eliminate corruption by making it harder for people to take up false or duplicate IDs to access entitlements and subsidies, especially among the poor, but it has seen many challenges on the ground when biometrics fail or the technology fails due to poor network connectivity, and poor people have been denied their entitlements. Data privacy and security is another key concern, with cases where the Aadhaar numbers of people have been leaked due to seemingly careless implementation, and allegations have even been made about security lapses of unauthorized access to biometric data and other personal details.

3.2.2 Cashless Economy

Cashless Economy [42] aims to create an economic state whereby financial transactions are not conducted with money in the form of physical banknotes or coins, but rather through the transfer of digital information between the transacting parties. Among other objectives, the demonetization move was positioned as a policy to push India towards a cashless state, so that the poor who do not have any credit history or access to banking channels, will be able to create this data trail that will help them later get easier access to formal sources of credit and other financial instruments. Controversy however prevails, because the low access to digital technology by the poor, trust issues in using entirely electronic means for money management, capability and skills at technology usage, and even the low utilization of bank accounts by the poor, lead to arguments about whether the country is even ready for such a move and what kind of safeguards in law and skilling should be developed before such policies can be pushed.

3.2.3 Digital India

Digital India is a campaign launched by the Government of India, which includes plans to connect rural areas with high-speed Internet networks. The underlying assumption is that easier and cheaper access to the Internet will lead to development of the poor, but many researchers have argued that such an infrastructure push should also be accompanied by a digital literacy campaign to alert first time users of information technology about problems caused due to undesirable appropriation of technology, as has been evidenced with a rise in rumors and fake news on social media and messaging platforms. The policy and the associated push for digital payments [31] can be argued to be an example of the high modernism approach often undertaken by the state in assuming a validity in its approach without any testing [33], and has similarly been noticed in another technology-driven initiatives like Aadhaar in their assumed validity without adequate testing [16].

3.2.4 E-Governance

The National e-Governance Plan (NeGP) is an initiative of the Government of India to make all government services available to the citizens of India via electronic media, instead of them having to fill up paper forms. However, an emphasis on electronic means of accessing government services at the cost of not scaling traditional offline face-to-face mechanisms, is known to cause problems to the poor because of their limited technological skills, technology access, and empowerment. Studies have pointed out how E-governance initiatives have been primarily shaped by the political economy around the policy, at times leading to its misuse. A study by Benjamin et al. [7] found that the e-governance project Bhoomi, which

digitizes land records around Bangalore in India, led to corruption and capture of large areas of land by elites with connections to the judiciary and administration, aided by the land-records data that became more easily accessible for strategic planning by these large players. The authors conclude that in policies related to E-governance, the outcomes are shaped more by the political economy around policies rather than the techno-managerial concerns. Veeraraghavan [41] in his work studied the data management platform built for NREGA, and found that the platform originally intended to curtail corruption and bring transparency, ended up mostly being useful as an accounting and reporting platform but was not able to address corruption due to new pathways discovered by local officials to bypass technology checks.

Chapter 4

DATA

We analyze policies and events of national importance, which have been widely discussed and debated in the mass media. Our analysis pipeline contains a number of steps similar to those described in Sen et al. [34]. We have built crawlers to collect mass media data on a daily basis from some of the most popular national news sources in English, *The Times of India, Indian Express, The New Indian Express, Telegraph, Deccan Herald* and *Hindustan Times*, and archives are used to build a corpus of news articles since 2011.

4.1 Collection of Articles

We have built a corpus of media articles published by Indian English-language news sources by crawling their web archives / RSS feeds. This system can be used for analyzing media articles, getting relevant articles by efficient querying or integrating the system with other data sources to mine interesting patterns. The media corpus that we build using the news article crawlers consists of articles belonging to the categories: National, International, Regional, Sports, Opinion and Business. This categorization is followed by all news sources in general. National, international and regional are political news articles. Some articles are also tagged as FRONT PAGE if this information is available while crawling their URLs. Sometimes URLs are published under multiple categories by news source. So we associate an array of categories for each article. This data is stored in MongoDB which is a No-SQL database. An article document in MongoDB consists of its meta data which is stored in Postgres along with text and author info. An article document in MongoDB has fields as shown in the table 4.1.

In order to extract a subset of articles relevant to a particular policy from this set, we filter the entire collection of articles with some domain-specific keywords. To identify news articles about a topic, we first supply a list of manually selected keywords corresponding to each policy. After extracting articles containing these keywords, the keyword set is expanded with newer keywords from these articles, based on their frequency. These two steps are repeated iteratively until the keyword set becomes static, and the final set of articles is used to perform our analysis. The final set of augmented keywords for each policy is shown in table 4.2.

For the economic domain, we perform our analysis on 22,302 articles on Demonetisation (Nov 2016 to Oct 2019), 13,908 articles on Aadhaar (2011 to 2019), 22,179 articles on GST

Table 4.1: Relational database schema of an article in MongoDB

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

(Jan 2011 to Oct 2019) and 85,486 articles of Farmers' Protests (Nov 2016 to Oct 2019). The periods for Demonetisation and GST were identified around the immediate months when the policies came into effect. Aadhaar and Farmers' Protests have had long standing debates, and therefore a longer period of time was used for these topics. Following a similar methodology, we were able to extract 23,432 articles (Jan 2014 to Oct 2019) relevant to technology policies.

4.2 Entity Extraction: OpenCalais

We then extract entities from this article data using the OpenCalais tool [1], which returns entities of type Person, Company, Organization, City, Province, and Country that are stored in a document database. OpenCalais extracts entities of many types but we use only a few of them in our system. Also, only a subset of properties of an entity are being used. We use entities of type Person, company, Organization, Country, City, Continent, ProvinceOrState and properties name, type, instances (entity references in article), resolutions (unique id given by OpenCalais from its database). Extracted entities are added to the article document stored in MongoDB. Each entity is then enriched with contextual information and stored in a separate entity collection. This collection might contain duplicates. Henceforth, we will refer to this collection as unresolved_entities. OpenCalais tries to resolve extracted entities by searching them in its database, but mostly gives accurate ids for places, somewhat accurate for companies and poor for persons. Along with the entities, OpenCalais also provides additional context attributes like the type of entity, its standard name, and some other context information (especially for the non-person entities like latitude and longitude for locations). We also maintain a set of aliases for each resolved entity, which keeps getting

Table 4.2: Policies and the set of augmented keywords to extract articles from mass media

Policy	**Keywords (Manually Selected)**	**# of articles**
Aadhaar	aadhar, aadhaar, UIDAI, adhar, adhar card, aadhar card, PDS, public distribution system	13,908
Demonetisation	demonitisation, demonitization, denomination note, cash withdrawal, swipe machine, unaccounted money, withdrawal limit, pos machine, fake currency, digital payment, cash transaction, cashless economy, black money, cash crunch, currency switch, long queue, demonetised note, cashless transaction, note ban, digital transaction	22,302
GST	gst, gabbar singh tax, goods service tax, goods and services tax	22,179
Farmers' Protests	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	85,486
Technology	privacy, cashless, technology, technological, innovation, software, engineering, smart city, technical, data protection, big data, artificial intelligence, digital india, high speed internet, make in india, e-governance, umang, digital literacy, national policy on electronics, e-gadget, entrepreneur, startup, scientific, science	23,432

enriched with standard names of the newer entities that are resolved with it. Currently, we are using English language news sources for our analysis, since Open Calais works only for English. As part of future work, we are also attempting to analyze news articles from vernacular regional media sources.

4.3 Entity Resolution: Elastic Search

Entities extracted from media articles are inserted into unresolved_entities collection. This collection contains duplicates entities and they need to be resolved. For this we create a resolved_entities (initially empty) collection which will have resolved entities. There are millions of entities in our system and comparing with each one during resolution will be time consuming. To do this efficiently, we use Elasticsearch engine. It is a highly scalable open-source full-text search engine, popularly used as an underlying engine in applications for efficiently storing and querying large amounts of data. So there are two copies of resolved_entities collection (MongoDB & Elasticsearch) in the system. To resolve an entity_x from unresolved_entities, we search for it in resolved_entities using Elasticsearch which

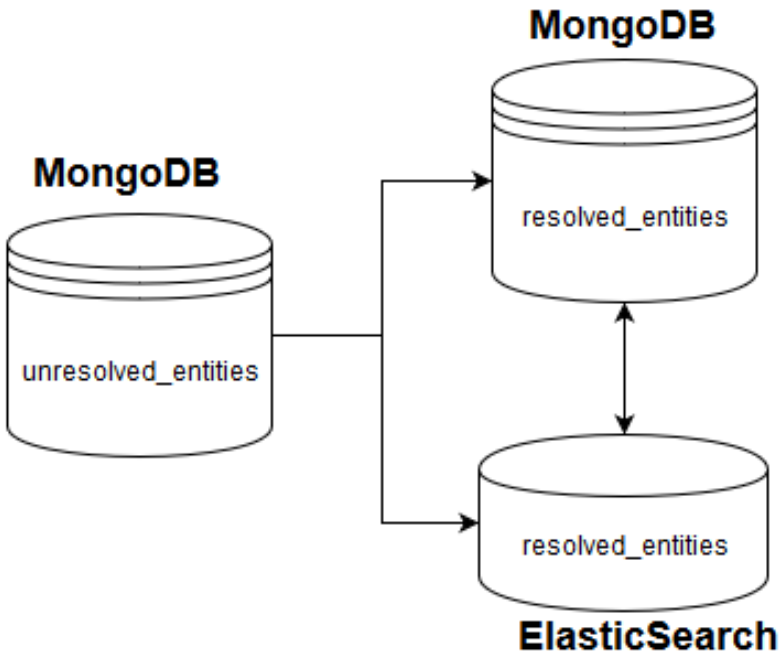


Figure 4.1: MongoDB and ElasticSearch Collections

gives ten best matched results. From these results, we find the best match for entity_x.

Since the same entity might occur in various forms (in terms of spellings and abbreviations) in different articles and news sources, we perform entity resolution (ER) within the media data as described in detail in [34]. 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 media data follows two steps:

1. It finds the top ten candidate entities from the resolve set based on partial matching of their standard names, aliases, and context
2. 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

[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.

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.

4.4 By-statement Extraction

To understand the ideological bias carried by news-sources, we need to extract the statements made by influential entities on the policy issues in these news-sources. These entities include politicians, business-persons, bureaucrats, social activists, and others. Statements on policies that occur in article text can be divided into three classes: the *by class* (containing statements made by the entities covered in media), the *about class* (statements made by the media house about the entities), and the *Others class* (statements that are neither spoken by the entities nor are about the entities). We perform entity resolution (ER) of these entities, which results in the identification of various aliases used to refer to an entity, along with various other entity-specific keywords. For example, the aliases found by the ER process for the current Prime Minister *Narendra Modi* are *Modi, NaMo, Modi Ji,* etc. and the keywords identified with him are *PM, Prime Minister, Gujarat CM* (since he is a former Chief Minister of Gujarat), etc.

For every article, we first split the article into meaningful single-line sentences. This preprocessing has been done using a careful regex matching that incorporates ways that handles many of the more painful edge cases that make sentence parsing non-trivial e.g. *"Mr. M.P. Sharma was born in the U.K. but earned his Ph.D. in India before joining Tata Steel Inc. as an engineer. He also worked at ukmotors.org as a business analyst."* Some rules that have been used for handling such cases have been mentioned in table 4.3.

Table 4.3: Rules for Statement Extraction from Articles

Case	**Example**
numerals	1.2 million | 3.54 crores | ..
alphabets	e.g. | m.s. | ms. | sr. | X.y
prefixes	Mr. | St. | Dr. | Rs.| Prof. | Capt. | Cpt. | Lt. | Mt.
suffixes	Inc. | Ltd. | Jr. | Sr. | Co.
mixtures	Sr. A.K. Anthony | Mr. M.P. Sharma | A.B. Ltd. Co.
acronyms	U.S.A | U.K. | U.P.S.C. | W.W.W.
websites	.com | .net | .org | .io | .gov | .me | .edu
special characters	Ph.D. | M.Tech. | B.Tech. | etc. "..." | "!" | "?" | "\"

For a given entity, using the entity-specific keywords and aliases, we can sample out the by-statements by following the Algorithm 1. We elaborate the algorithm here in some detail: firstly, for every article associated with a given entity name, all occurrences (or aliases) of

Algorithm 1: *By-statement* Classification of a given text

// **Text**: Text statement obtained after splitting
// **EntName**: Entity Name whose *by-statements* to extract
// **Keywords**: List of entity-specific keywords
// **Aliases**: List of different aliases used for a given entity
// **FixedWords**: ["says", "said", "asked", "told", "claimed", etc.]
Input: *Text*, *EntName*, *Keywords*, *Aliases*, *FixedWords*
Output: *Classified Label* (By/About/Others)

```
shortName ← Short Entity Name for EntName
Replace all Aliases occurences in Text with shortName
pText ← Dependency Parse Tree of Text
Tags ← Identify noun chunks (nsubj, dobj) in pText
posText ← POS Tagging of Text
for pt in pText do
    check1 ← if pt is related by one of the Tags
    check2 ← if corresponding posText entry is in Keywords
  if check1 and check2 then
      check3 ← if pt is a part of an 'nsubj' relation
      check4 ← if corresponding pText is in FixedWords
    if check3 and check4 then
        return "By"
      else
        return "About"
return "Others"
```

the entity in the article are replaced by short, entity-specific tags (e.g., the aliases of the entity *Narendra Modi* can be *Modi*, *Modi ji*, *Namo*, etc. which are then replaced by the tag "narendra-modi" in the article). We then use Stanford CoreNLP [21] to obtain the parts-of-speech (POS) tags and dependency parse tree of the statement made by that entity. Sen et al. [34] have shown that several key dependency relations like 'nsubj', 'nmod', 'amod', etc. can be used to identify the class to which a statement belongs (by, about, or others). We also check if the object or the subject in the statement is connected to keywords like 'said', 'claimed', 'announced', 'told', etc. in the parse tree, which indicates that the statement belongs to the *by* class. These *by-statements* are then finally returned by the algorithm.

4.5 Relevance Filtering

Since we use a keyword-based approach to extract articles on a policy, we find that some statements in these articles do not talk about the policy, despite containing a keyword relevant to it. These statements can potentially hurt the accuracy of the classifier and need to be removed from our final dataset used for ideology classification. For example, with respect to the Aadhaar policy, *"Aadhaar project is an example of using modern technology to leapfrog*

Table 4.4: *By-statement* Distribution for various policies for the ground truth data after manual annotation

Domain	Policy	Total Statements	Relevant Statements				
			Pro	Anti	Neutral	Balanced	Total
Economic	Aadhaar	392	169	34	146	1	350
	Demonetisation	1255	512	259	243	49	1063
	GST	681	292	145	167	46	650
	Farmers' Protests	1899	961	505	262	64	1792
Technology		1075	553	115	134	10	812

for future development and transformation of a country, UIDAI chairman Nandan Nilekani has said".” is a relevant by-statement. On the other hand, 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 irrelevant, since it does not discuss about any policy in particular and talks about the winning candidate in a poll. It, however, gets extracted since the keyword *Aadhaar* appears in it. We use an annotated dataset (ground truth) for the relevance-based filtering process, which has been created through manual inspection. 10.7%, 15.3%, 4.55%, and 5.63% of the statements in Aadhaar, Demonetisation, GST, and Farmers' Protest were found to be irrelevant, respectively. Similarly, there were 24.4% irrelevant statements for the Technology policies. The distribution of relevant statements can also be visualised from figure 4.2. We now elaborate the different methods to filter out the irrelevant statements, and compare their accuracies.

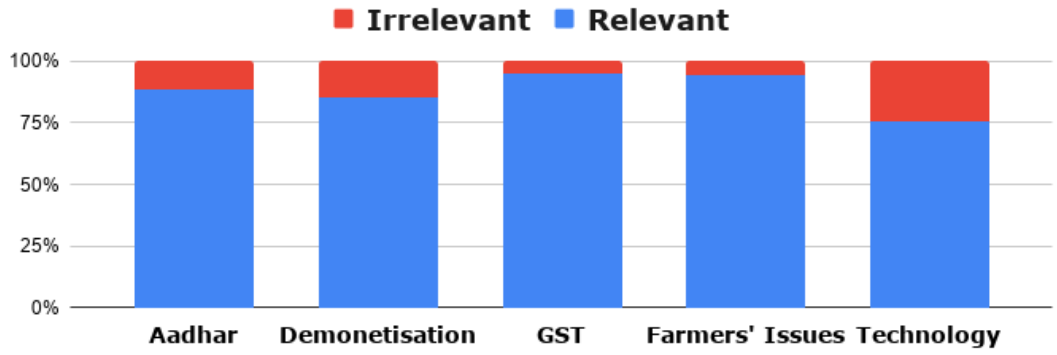


Figure 4.2: Distribution of Relevant and Irrelevant Statements

4.5.1 Rule-Based Approaches

Analysing the dependency parse tree of a by-statement helps in identifying important relationships between various subject (e.g. ‘nsubj’, ‘csubj’, ‘nsubjpass’, etc.) and object tags (e.g. ‘dobj’, ‘iobj’, ‘pobj’). The dependency ‘nsubj’ shows the relationship between the main predicate and the subject, while the ‘dobj’ dependency links the predicate with the

object. The 'xcomp' dependency gives internal information about the predicate. After preliminary analysis of various dependency graphs, we found some commonalities in the structure of most of the relevant statements. These relevant statements usually are related by the 'nsubj'/'dobj'/'pobj' dependencies, and contain specific domain-related keywords in the noun-chunks (base noun-phrases). However, these noun-phrases need not be directly tagged as the 'nsubj'/'dobj'. A sample statement: *"At UIDAI, we are concerned about the privacy issue."* has been parsed using the built-in dependency visualizer of spaCy in figure 4.3. Some of the important entities related by some object and subject tags have been highlighted in red.

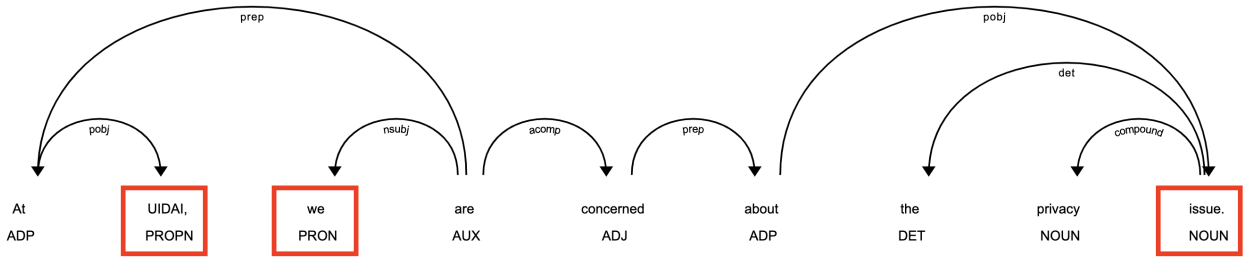
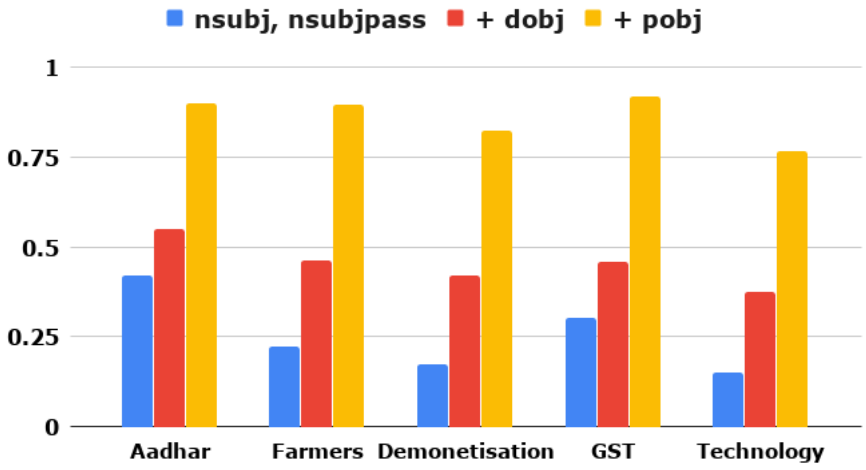


Figure 4.3: Dependency Parse Tree for a By-Statement

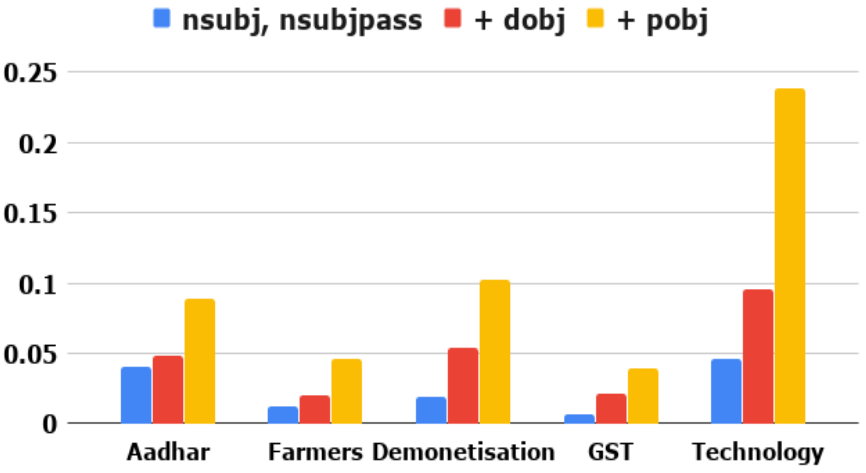
Since our relevance filter is a check before the main classification stage, we are majorly concerned with high precision in the relevant statement detection process. Thus, an ideal check should have some high value for true positives and a relatively lower value of false positives. On increasing the number of tags, we see an increase in the number of true positives as seen in figure 4.4a. But, from figure 4.4b, we also see that increasing the number of tags also increases the number of false positives. Hence, we choose the set of object and subject tags so as to find a right balance between the two. We also measure the F1-score for each policy and draw its performance comparison with other method in table 4.5.

Table 4.5: F1-Scores for Relevance Check by Rule-based and Supervised methods

Policy	**Rule Based Approach**			**Supervised**	
	nsubj	***+ dobj***	***+ pobj***	**Random Forest**	**Gradient Boosting**
Aadhaar	0.58	0.69	0.91	**0.95**	0.91
Demonetisation	0.29	0.57	0.85	**0.92**	0.85
GST	0.46	0.62	0.94	**0.98**	0.97
Farmers' Protests	0.36	0.62	**0.93**	0.92	0.84
Technology	0.25	0.51	0.77	**0.80**	0.75



(a) True Positives



(b) False Positives

Figure 4.4: Relevance Check by Rule-Based Approach

4.5.2 Supervised Approaches

H. C. Wu et al. [46] has shown the significance of TF-IDF (Term frequency-inverse document frequency) term weights in making "document-wise" relevance decisions. Following upon the idea, we convert our by-statements into a similar TF-IDF vector representations for building a supervised approach. We perform a binary classification using various non-neural machine learning algorithms like SVM, Random Forests, Gradient Boosting, etc. Our corpus consists of equal number of relevant and irrelevant statements. From multiple policy distributions, we find that Random Forest works the best in most of the cases, in the supervised scenario. Our features have a good predictive power, and the randomness across the features coupled with the bootstrapping in Random Forest gives rise to a set of uncorrelated predictions in the trees of the forests. This boosts up the accuracy by a large margin. This is also highlighted from our results in table 4.6.

4.5.3 Unsupervised & Semi-Supervised Approaches

To reduce the dependence on the labels, we experiment with some unsupervised and semi-supervised approaches as well. For each collection of by-statement per policy, we can use Latent Dirichlet Allocation (LDA) [8] to automatically discover the topics that these sentences contain. LDA is a bag-of-words model that assumes that documents (sentences) are a mixture of topics produced through a generative process. For each policy set, we can safely assume that there are 2 topics - one relevant to the policy and the other irrelevant to the policy. We run LDA with unigram and bigram features for each stack of statements corresponding to each policy with a topic count of 2. From results in table 4.5, we see how LDA is helpful in determining the topic of relevance in the corpus, but fails to accurately model the non-relevant categories because of the ambitiousness inherent to it.

Table 4.6: F1-Scores for Relevance Check by Unsupervised and Semi-Supervised methods

Policy	**Unsupervised**			**Semi-Supervised**
	Kmeans		**LDA**	**Guided LDA**
	Count	**TF-IDF**		
Aadhaar	0.81	0.82	0.76	0.83
Demonetisation	0.85	0.83	0.72	0.71
GST	0.84	0.81	0.66	0.69
Farmers' Protests	0.79	0.73	0.62	0.68
Technology	0.76	0.77	0.69	0.70

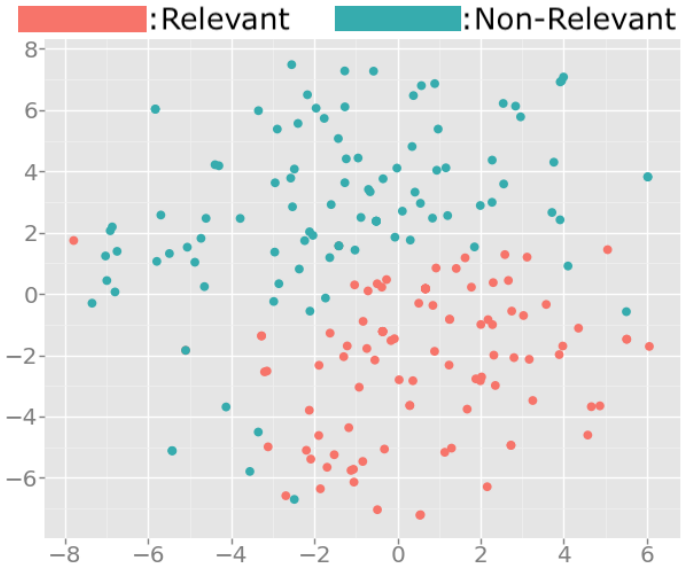


Figure 4.5: t-SNE plot of Count-Vectorizer Embeddings

On the other hand, we also experimented with K-Means[20] clustering to visualise the shape of the decision boundary for relevant and irrelevant labels. We find that both the classes are separable under a simple count vectorizer embeddings as seen in figure 4.5. We

also exploit the idea of Guided LDA [15] to set some seed words for relevant and non-relevant categories for each policy, which can guide the model to converge around those terms. Guided LDA performs slightly better than Seedless LDA which is conclusive from the fact that Guided LDA provides a direction (seed) to mitigate the ambiguity present in the non-relevant topics. tSNE for Count Vectorizer is shown in figure 4.5. F1-Scores for the relative comparison between Count & TF-IDF vectors have also been shown comprehensively in table 4.6.

Conclusion: Since RF turns out to be the most robust in terms of F1-Scores among all these methods, we choose to deploy an RF-based pre-trained model for checking relevance. We are able to achieve 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, using RF. Performance comparison between these methods in terms of F1-scores has been described in table 4.5 and 4.6.

Chapter 5

RESEARCH PROBLEM

5.1 Problem

Different Classes of By-Statements

A statement can be classified into different categories based on the political ideology it holds. A *Pro* statement is where the speaker is in support of the policy or appreciates the policy (e.g. *"Aadhaar project is an example of using modern technology to leapfrog for future development."*). *Anti* statements are where the speaker criticizes the policy or talks about its drawbacks (e.g., *"Despite waiver, banks have still not started disbursing fresh credit to farmers leaving them starved."*). *Neutral* statements do not have a specific stance (e.g. *"Speaking at the opening ceremony, Shukla said, the BCCI is committed to the welfare of farmers."*). There also are some statements that are diplomatic in nature and hold both *Pro* and *Anti* stances. We collectively group such statements into another class called the *Balanced* statements. The Pro/Anti Econ classifier takes as input an economic policy related statement made by an entity and outputs whether it is in favor of (pro) or against (anti) that policy.

A technology related statement can be categorized under different groups depending on its ideological position. *Pro* technology (Technology Determinism) statements are where the subject shows faith in technology [45], and often suggests technology as the solution to people's problems (e.g., *"Mr. Modi told media leaders that digital technology can help in innovation and empowerment."*). *Anti* technology (Technology Skepticism) statements show doubt and skepticism about using technology, or the problems with its implementation. (e.g., *"Matching such state-of-the-art systems could be a technological nightmare for Indian counterparts."*). The Tech Determinism/Skepticism classifier takes as input a technological intervention related statement made by an entity and outputs whether it shows faith (determinism) or doubt (skepticism) towards that intervention.

We classify these *by-statements* into different classes (ideological positions) depending upon the policy domain (Economic/Technology) in which they lie. Some example of *by-statements* have been mentioned in table 5.1. Further detailed examples of each class of *By-statements* have been mentioned in the appendix A (tables A.1-A.1).

Goal

We wish to study the ideological bias carried by Indian mass media in terms of Pro, Anti, and Neutral statements on important economic and technological policies. In this direction,

Table 5.1: Examples of different classes of Relevant *by-statements* extracted

Class	**Policy**	**By-Statements**
Pro	Technology	*Mr. Modi told media leaders that digital technology can help in innovation and empowerment.*
	Aadhaar	*Aadhaar project is an example of using modern technology to leapfrog for future development.*
Anti	Technology	*Today, we misuse technology and kill girls in the womb of the mother, Modi said, adding that the damage being done through generations would take another two to three generations to be rectified.*
	Farmers' Protests	*Despite waiver, banks have still not started disbursing fresh credit to farmers leaving them starved.*
Neutral	GST	*There is no centralised exemption but if a State wants, it can refund the SGST for regional promotion.*
	Farmers' Protests	*Speaking at the opening ceremony, Shukla said, the BCCI is committed to the welfare of farmers.*
Balanced	Technology	*Taking on the opposition and accusing it of misguiding people over cashless transactions, Modi said that while on one hand they claimed that Rajiv brought about the telecom revolution, on the other they\ \contradict themselves by saying that their countrymen don't have means for cashless transactions.*

we aim to use natural language processing with deep learning to build a classifier that can detect the stance (neutral/non-neutral) and classify the ideological position (Pro/Anti) held by influential entities that make statements on policies in the mass media. We also target to keep our methodology generalizable so that our models can also be applied to any other policy domain. More specifically, we enlist the two key problems that we target:

1. Our aim is to build a general pro/anti economic policy classifier, which does not require a dataset for each new policy for which it performs classification. In other words, we want to build a model that trains on a subset of policies and learns the various common axes of economic debates, and then classifies the statements belonging to any other economic policy on these same axes.

2. Similarly, our goal is to create a general pro/anti technology classifier to evaluate the degree of tech determinism or skepticism prevalent in our mass media sources. We aim to build the model using a small subset of policies concerning the technological domain, from which it can learn the pattern to classify any new interventions in the tech-policy area accurately. The dataset for this model is entirely different from the econ classifier in the previous part.

5.2 Dataset Annotation

With the help of our research group, we first manually code the *by-statements* to one of the five classes - *Non-Relevant, Pro, Anti, Neutral*, and *Balanced* using a coding schema that we describe further in thisw section. We resolve ambiguous cases using the context information that we preserve while extracting the *by-statements*. The context information includes the preceding and succeeding statements corresponding to each *by-statement* in the format ⟨*preceding-statement*, *by-statement*, *succeeding-statement*⟩. For example, for the sentence, *"The technology has undergone a drastic transformation in the last 20 years, Modi said adding the aspirations of the youths have to be kept in mind in this era"*, the following is the context information being stored using the script: (*"The PM said India's economy is being transformed and the manufacturing sector is getting a boost."*, *"The technology has undergone a drastic transformation in the last 20 years, Modi said adding the aspirations of the youths have to be kept in mind in this era."*, *"On merits of democracy, the Prime Minister said, Bigger than the strength of the government is the people's power."*). Here, if we consider the words, 'boost' in the preceding statement and 'the effect of technology' on the economy and manufacturing sector, the statement appears to be Pro-Technology.

5.3 Coding Schema

We construct a coding schema that helps the annotators code the statements to one of the five classes (Non-Relevant, Pro, Anti, Neutral & Balanced). The use of a coding schema to unambiguously code data is quite widespread in the field of qualitative content analysis [5, 39]. In our case, the goal of the schema is to guide an annotator to understand if a *by-statement* conveys an ideology about a policy. The format used for coding schema is described in table 5.2 for reference. For each policy, the schema contains the possible targets of a policy statement and the frequently occurring keywords with suitable examples for each class label. It helps the annotators accurately code any new statement on a policy. For example, statements that convey measures to bring relief from problems faced during Demonetisation or statements that tend to show the advantages of this policy in curbing corruption and black money, convey a pro-policy ideology. On the other hand, statements that highlight the failure of the policy and difficulties faced by the people because of having to stand in long queues, or due to shortage of cash, hold an anti-policy ideology. To ensure consistency across different policies, these definitions are suitably customized for each event. Thus, for Farmers' Protests, articles referring to provision of loan waivers and availing crop insurance tend to protect the interests of the farmers and are designated to be pro-policy in nature. We also provide relevant keywords and suitable examples for each class label particular to a policy in the schema.

A normative definition of each class is also provided to help the annotator understand the general intuition of the class before labeling. The coding schema has been built after manually studying roughly 100 articles from each policy event (a manageable size of 400 articles in total) by the lead authors of this study, assisted by two annotators. After multiple rounds of due deliberation to reduce subjectivity, the coding schema is finalized and given to the rest of the annotators to perform the final labeling. The inter-coder reliability (using Cohen's Kappa statistic [23]) of the labeling exercise, which was evaluated by taking a random sample of 100 statements from different policies by three annotators pairwise, comes out to be roughly 0.75-0.79. It means that our gold set of labels is reliable and robust enough for various analytical experiments. Our coding schema for all the policy events can be found in the appendix A (figures A.1 - A.4 and figure A.5).

Table 5.2: Format of the coding schema (for a policy) used for annotations

Column	**Description**
label	Annotation of the class
keywords	Keywords representing the class
examples	Some examples of the class to understand the label
normative definition	A more general intuition of the class label
exceptions	Some corner cases of labeling
annotatorś feedback	Feedback given by the annotator regarding the schema

Chapter 6

METHODOLOGY

In this section, we describe our framework to detect the political ideology of a particular *by-statement*. We start off by describing the model of the classifier in detail and then explain the procedure followed for training for making it a generalizable, entity-independent classifier.

6.1 Model

We use a two-step classifier to predict whether a statement has *Pro*, *Anti* or *Neutral* alignment for a policy. The first step – *stance detection* – is to determine whether the statement holds any stance concerning the policy or not, i.e., whether it is neutral or non-neutral. The second step – *ideology classification* – aims at classifying whether the statement is in support of the policy or against it, in case it is non-neutral. Models built for both of these steps have the same architecture and training method.

We use Recursive Neural Networks (RNN) inspired by Iyyer et al. [14], as our predictive model. They are a type of hierarchical neural network which take into account both the syntactic and semantic features of the sentence and have been known to achieve state-of-the-art performance on various NLP tasks like sentiment classification, parsing, paraphrase detection and political ideology detection. They work on the assumption that the meaning of each phrase should be a combination of the meaning of the words that form it and the syntax that combines these words. Each phrase of a sentence can represent different ideologies, which combine to reflect the overall ideology portrayed by a sentence. The structure of the RNN takes this into account by first predicting ideologies of the phrases at a low level and then combining them with learned weights in a bottom-up manner to predict the ideology of the overall sentence. This kind of a network takes into account the structure of a sentence as well as the meanings of its individual words. This makes it effective as compared to other approaches that are based on the absence or presence of certain words or phrases in the sentence. It is also more effective than the approaches which use HMM-based models, since it combines information in a hierarchical rather than a linear manner, thus building the meaning of a sentence from its relevant phrases. Since it was not feasible to annotate each phrase of a sentence, we assume that the label of a phrase is the same as that of the *by-statement* in which it occurs. To break the sentence into phrases, we use the Stanford Parser to obtain the parse tree of the sentence, which is fed to the model as the input. For example, for the sentence, *"Digital payments will improve the functioning of toll plazas, Das said."*, the PoS text and the parse tree can be found in the figure 6.1.

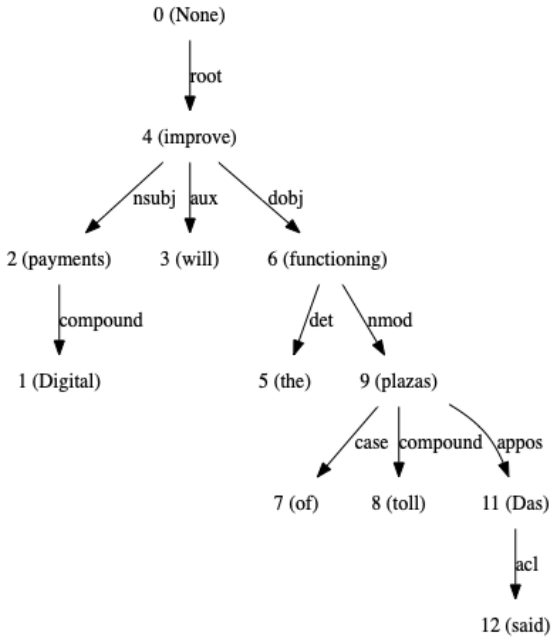


Figure 6.1: Example of a Parse Tree

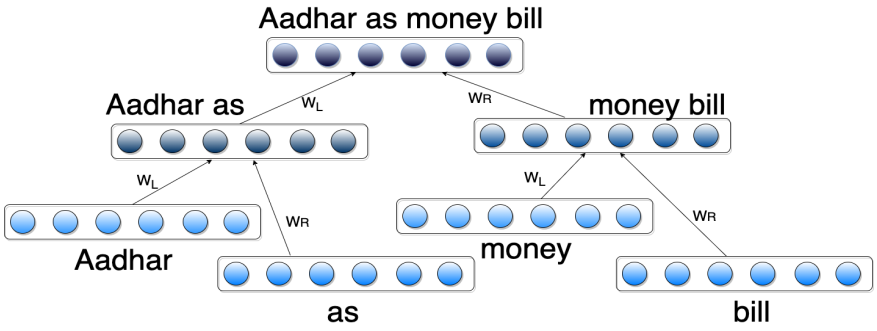


Figure 6.2: Example of how word representations are combined to form phrase vectors of the same dimensions

To represent the phrases of the parsed sentence as vectors, embeddings of words forming a particular phrase are combined to form a *phrase vector* (figure 6.2) that 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 according to the following equation:

$$x_p = f(W_L.x_a + W_R.x_b + b_1)$$

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 of dimension $d \times d$, and b_1 is a bias term of dimension $d \times 1$. 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 of dimension $2 \times d, 2 \times 1$, and the softmax is

$$\text{softmax}(q) = \frac{\exp(q)}{\sum_{i=1}^{k} \exp(q_i)}$$

We use a cross-entropy loss function for training, which for a single statement is given by combining the loss for all the phrases of the sentence like

$$l(\hat{y}_s) = \sum_{p=1}^{k} y_p * \log(\hat{y}_s)$$

The final cost function consists of a sum of the losses over all the statements in the corpus. Because of a small-sized dataset, we use $L2$ regularisation to avoid overfitting. The parameters of the model are optimized using Stochastic Gradient Descent with momentum.

To estimate the performance of the overall model, we employ macro-averaged F-score $F_{macro} = \frac{F_{Neutral}+F_{Non-Neutral}}{2}$ at Step-1 (stance classification) and $F_{macro} = \frac{F_{Pro}+F_{Anti}}{2}$ at Step-2 (ideology classification).

6.1.1 Word Representation

We use the embedding matrix from word2vec [24] pre-trained on the Google News corpus to initialize the embedding layer of the model. We have two choices while training the model: to train the embedding layer along with the whole model or to freeze this layer while training. We choose to do the latter as we do not have enough training data to learn good representations of words while also ensuring a model with good performance. The results in chapter 7 bolster this claim by showing a better model performance with a frozen embedding layer. While freezing this layer, we rely on our assumption that the initial embedding layer matrix gives a good enough representation of the word, requiring no additional training. However, as also shown in chapter 7, simply using the pre-trained vectors does not perform well, since ideology detection is a complex task that requires fine-grained embeddings, which are aware of the political context in which a word is used. For example, words like "digital" and "smart" have a very fine-grained meaning in this domain as they are more likely to refer to the policy interventions like *"Digital India"* and *"Smart Cities"*. Hence, we take the pre-trained Google News embeddings and fine-tune them before feeding them to the classifier without using any additional supervision. This is done by re-training a word2vec model, initialised with the pre-trained embeddings, on domain-specific policy corpus. We use two collections of articles, one about economic policies and another about technology policies, to fine-tune the word embeddings for economic and technology classifiers, respectively, which are then used to initialize the first layer of the model.

6.1.2 Making classification entity-independent

There is a reasonable improvement in the performance of the classifier when fine-tuned embeddings are used. To understand its effects on the performance, we also look at some examples which are being misclassified while using fine-tuned embeddings but are correctly classified when using general-purpose embeddings. We find that fine-tuned embeddings of entities exhibit associations with certain words. For instance, words supporting a policy have more association with ruling party entities (who rolled out the policy) compared to opposition party entities. We also find that dominantly anti-policy words associate significantly with entities of a specific religious group. Such associations often mislead the classifier into almost always favoring the ideology held by the entities while ignoring the hidden semantics of their statements. The process also captures undesirable associations such as caste and religious stereotypes.

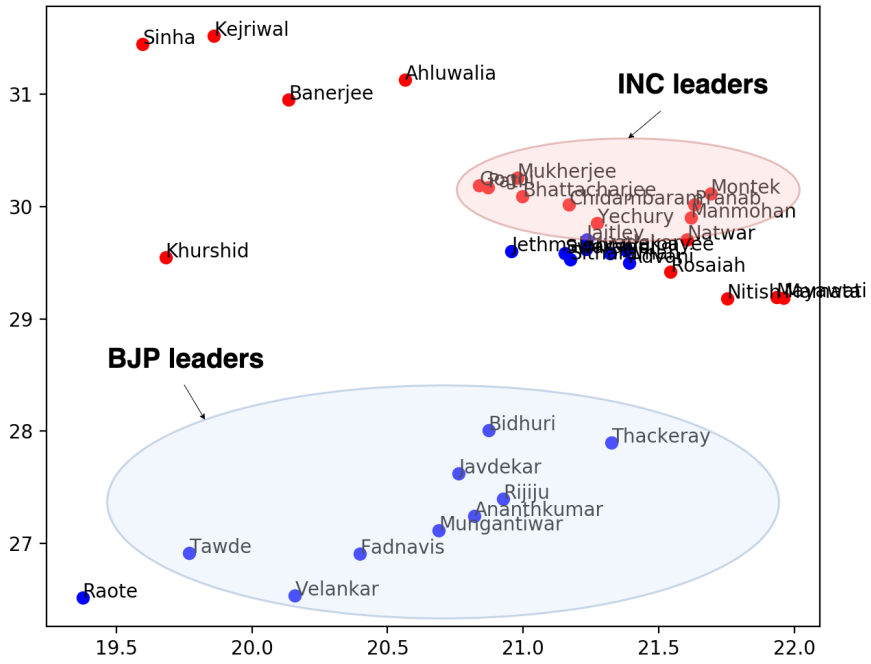


Figure 6.3: Leaders in support (blue) & opposition (red) of economic policies

Figure 6.3 qualitatively shows some of the associations by visualizing the representations through a t-SNE plot. As shown, leaders in support (mostly of the ruling party, BJP) and in opposition (mostly of the opposition party, INC) of the policy form separate clusters, which explains the dependence of classifier predictions on entities present in a sentence. Further, leaders of BJP form a separate cluster, which is closer (measured using cosine-similarity) to clusters of words like "anti-corruption", "cashless" which are associated to *Pro* stance concerning economic policies launched by BJP. To mitigate this dependence, we compare two methods of removing the named entities from the corpus before fine-tuning – remove these entities (black them out), or replace them with their named entity tags. From section 7, we find that fine-tuning results in better overall performance after replacing the named

entities with their tags, since it preserves the grammar and structure of the sentence. The proposed model is hereinafter referred as **ID-RNN** (Ideology Detection - Recursive Neural Network).

Chapter 7

RESULTS

To account for class imbalance in the dataset, all the experiments are done on a balanced dataset by either undersampling the majority class or oversampling the minority class.

Table 7.1: Baseline Performance of Non-Neural Methods on Stance Detection (**U**: Undersampling, **O**: Oversampling, **Acc**: Accuracy %, **F1**: F1-Score)

Models	**Step-1 (Stance Detection) Neutral vs Non-Neutral**			
	Economic		**Technology**	
	U	**O**	**U**	**O**
	Acc% (F1)	**Acc% (F1)**	**Acc% (F1)**	**Acc% (F1)**
SVM	66.3 (0.63)	74.1 (0.61)	57.4 (0.50)	78.3 (0.54)
GB	62.6 (0.58)	65.4 (0.60)	40.8 (0.39)	61.9 (0.49)
DTs	59.8 (0.55)	69.3 (0.58)	45.8 (0.44)	66.7 (0.52)
RFs	65.9 (0.61)	**78.2 (0.63)**	41.8 (0.40)	**79.1 (0.52)**
Rocchio	68.1 (0.62)	69.7 (0.67)	**62.8 (0.56)**	78.1 (0.59)
kNN	52.9 (0.51)	66.5 (0.60)	43.8 (0.42)	72.4 (0.58)
NB	**70.3 (0.65)**	73.2 (0.65)	55.7 (0.48)	73.4 (0.61)
BoW	49.7 (0.49)	62.0 (0.58)	35.4 (0.35)	71.2 (0.56)

Table 7.2: Baseline Performance of Non-Neural Methods on Ideology Classification (**U**: Undersampling, **O**: Oversampling, **Acc**: Accuracy %, **F1**: F1-Score)

Models	**Step-2 (Ideology Classification) Pro vs Anti**			
	Economic		**Technology**	
	U	**O**	**U**	**O**
	Acc% (F1)	**Acc% (F1)**	**Acc% (F1)**	**Acc% (F1)**
SVM	75.8 (0.74)	70.8 (0.65)	75.4 (0.66)	**86.4 (0.36)**
GB	68.3 (0.67)	67.5 (0.63)	74.2 (0.66)	79.1 (0.43)
DTs	65.1 (0.64)	67.8 (0.63)	68.1 (0.73)	74.8 (0.38)
RFs	**75.9 (0.74)**	**75.3 (0.68)**	76.4 (0.69)	85.8 (0.62)
Rocchio	71.7 (0.70)	70.7 (0.67)	74.3 (0.65)	84.2 (0.55)
kNN	71.2 (0.70)	60.8 (0.64)	56.8 (0.53)	58.6 (0.35)
NB	73.3 (0.72)	73.4 (0.70)	**78.3 (0.67)**	85.6 (0.45)
BoW	68.2 (0.67)	39.8 (0.33)	63.2 (0.58)	39.8 (0.30)

7.1 Baselines

We have developed strong baselines for both Step-1 (Stance Detection) and Step-2 (Ideology Classification) based on machine learning and deep learning models in order to draw a comparison with our proposed model.

1. *Machine learning baselines:* We use the following machine learning algorithms as baselines: Linear SVMs, Gradient Boosting (GB), Decision Trees (DTs), Random Forests (RFs), Nearest Centroid (Rocchio), k-Nearest Neighbors Classification (kNN), Naive Bayes (NB), and Bag of Words (BoW). For each algorithm, we first create a vocabulary set on the training data and then convert the sentences into TF-IDF representation before classification. The best baseline for Step-2 (Ideology Classification) gives an accuracy of 75.3% (F1-Score - 0.70) for Economic Policies, and 86.4% (F1-Score - 0.36) for Technology Policies (with oversampling). The detailed results for these baselines for both the steps can be found in the tables 7.1 and 7.2.

2. *Deep Learning baselines:* Apart from a simple Deep Neural Network (DNN), we also experimented with the widely used Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs) [17] and Recurrent Convolutional Neural Networks (RCNNs) [19] as baselines. All these models are initialized with fine-tuned word2vec embeddings (except for DNN) and trained with 200 epochs using the Adam optimizer with a batch size of 32. In case of CNN, RNN and RCNN, we preprocess the dataset by first padding (append a <PAD> token) each sentence to the maximum sentence length of 59, and then, constructing a vocabulary index and mapping each word to an integer between 0 and 18,765 (the vocabulary size). Each sentence thus becomes a vector of integers. We now describe the architecture of these models briefly.
   (a) **DNN**: We use TF-IDF vectors as inputs to the DNN, which consists of 4 hidden layers with 512 nodes in each layer, with a 50% dropout following each dense layer.
   (b) **CNN**: Convolutional Neural Networks (CNNs) are generally used in computer vision, however they've recently been applied to various NLP tasks [17] and the results were promising. The result of each convolution will fire when a special pattern is detected. These patterns could be n-gram word expressions like *I hate*,*very good*,etc. and CNNs can identify them in the sentence regardless of their position by varying kernel size. It can also learn important words or phrases through selection from a max pooling layer. However, processing text is difficult with CNNs because learning an optimal kernel size is challenging. The first layer uses low-dimensional word vectors as input in the CNN, while the next layer performs convolutions over these word vectors using multiple filter sizes. Next, we max-pool the result of the convolutional layer to form a one-dimensional feature vector, add dropout regularization (of 50%), and classify the result using a softmax layer.
   (c) **RNN**: RNNs can exhibit dynamic temporal behavior for a time sequence. RNNs capture contextual information by maintaining a state of all previous inputs. Since RNNs favor more recent inputs, they're considered to be relatively more biased. This baseline model has 3 hidden layers with 256 GRU nodes in each layer and a recurrent dropout of 20%.

(d) **RCNN**: Recurrent Convolutional Neural Networks (RCNN) [19] is also used for text classification. RCNNs capture the contextual information with the recurrent structure to construct the text representation using a series of convolutions. It combines RNN and CNN and exploits the advantages of both techniques simultaneously. We use 1D convolutions (with ReLU activation) and 1D max pooling to obtain a good representation. This is followed by 4 hidden layers of 256 LSTM nodes with a recurrent dropout of 25%. The output is finally classified using a softmax layer after passing through a fully connected layer.

The results comparing the performance of our model with these baselines have been shown in table 7.3 and 7.4.

Table 7.3: Performance comparison of our model (Stance Detection) (**ID-RNN**) with DL baselines (**U**: Undersampling, **O**: Oversampling, **Acc**: Accuracy, **F1**: F1-Score)

Models	**Step-1 (Stance Detection) Neutral vs Non-Neutral**			
	Economic		**Technology**	
	U	**O**	**U**	**O**
	Acc% (F1)	**Acc% (F1)**	**Acc% (F1)**	**Acc% (F1)**
DNN	70.2 (0.55)	75.3 (0.41)	68.4 (0.66)	75.6 (0.53)
RNN	74.4 (0.44)	74.5 (0.58)	77.8 (0.55)	77.5 (0.52)
CNN	70.7 (0.57)	77.3 (0.63)	84.1 (0.70)	83.3 (0.61)
RCNN	71.4 (0.65)	76.1 (0.64)	81.8 (0.71)	80.8 (0.57)
ID-RNN	**78.1 (0.75)**	**78.5 (0.77)**	**84.6 (0.80)**	**86.7 (0.90)**

Table 7.4: Performance comparison of our model (Ideology Classification) (**ID-RNN**) with DL baselines (**U**: Undersampling, **O**: Oversampling, **Acc**: Accuracy, **F1**: F1-Score)

Models	**Step-2 (Ideology Classification) Pro vs Anti**			
	Economic		**Technology**	
	U	**O**	**U**	**O**
	Acc% (F1)	**Acc% (F1)**	**Acc% (F1)**	**Acc% (F1)**
DNN	75.1 (0.71)	74.5 (0.63)	74.8 (0.60)	86.3 (0.67)
RNN	72.4 (0.70)	74.2 (0.61)	79.0 (0.64)	84.5 (0.69)
CNN	74.2 (0.70)	75.0 (0.63)	83.3 (0.64)	87.9 (0.71)
RCNN	75.2 (0.72)	77.6 (0.74)	83.2 (0.65)	89.1 (0.78)
ID-RNN	**78.9 (0.79)**	**80.1 (0.81)**	**85.5 (0.82)**	**90.7 (0.93)**

7.2 Proposed Model (ID-RNN)

In this section, we describe how using different word embeddings and training procedures affect the performance of our model. Here, we don't explain the details about the compar-

isons and experiments for Step-1 (Stance Detection), but primarily focus on Step-2 (Ideology Classification) that also has similar results. As described in Section 6, word2vec has been used with various modifications in our model: (a) G-W: Generic embeddings obtained after training word2vec on the Google News Corpus, (b) D-W: Embeddings obtained after fine-tuning on a collection of news articles about economic and technology policies, (c) D-W-NER: Fine-tuned embeddings obtained using the policy corpus after replacing the *Person* and *Organization* type of named entities with their named entity tags, and (d) D-W-BOE: Fine-tuned embeddings using the policy corpus, after removing entities.

Table 7.5: Effect of different word vector initialization

Model	**Economic**		**Technology**	
	Accuracy	**F1**	**Accuracy**	**F1**
G-W	69.8%	0.76	80.2%	0.81
D-W	74.1%	0.79	84.8%	0.89
D-W-NER	**80.1%**	**0.81**	**90.7%**	**0.93**
D-W-BOE	71.6%	0.78	82.8%	0.71

Table 7.6: Effect of freezing embedding layer (**F**: frozen, **T**: trainable)

Model	**Economic**		**Technology**	
	Accuracy	**F1**	**Accuracy**	**F1**
G-W(F)	69.8%	0.76	80.2%	0.81
G-W(T)	65.8%	0.70	76.2%	0.76
D-W-NER(F)	**80.1%**	**0.81**	**90.7%**	**0.93**
D-W-NER(T)	73.7%	0.76	85.6%	0.88

Apart from using different types of embeddings, we also train the model using two methods: by allowing the input layer weights to train, and by keeping them frozen. For both the datasets, the minority class is oversampled to account for class imbalance. Initializing the embedding layer with D-W-NER, with a frozen embedding layer, gives the best results, with accuracies of 80.1% (F1-score - 0.81) for Economic policies, and 90.7% (F1-score - 0.93) for Technology policies. This is a significant improvement when compared to G-W with a trainable embedding layer, which gives accuracies of 65.8% (F1-score - 0.70) on Economic policies, and 76.2% (F1-score - 0.76) on Technology policies. This boost in performance is explained in section 6.1.1. The results demonstrating the effect of different word vector initializations and keeping the embedding layer trainable/non-trainable can be found in table 7.5 and table 7.6 respectively.

We also observe that using a cascading two-label classifier (Pro/Anti) rather than a three-label (Pro/Anti/Neutral) classifier provides better results, with the latter resulting in F1-scores of 0.69 and 0.84, as compared to 0.81 and 0.93 of the former on the economic and technology datasets, respectively.

Chapter 8

Analysis

Several studies discuss how mass media is ideologically biased [37]. Sen et al. [34] show how the Indian mass media provides significant coverage to certain ideologies, rather than presenting a critical examination of policies. Budak et al. [9] talk about how the ideological bias in news organizations indirectly favors a particular side by criticizing the other side disproportionately. Our work is an improvement over the study by Sen et al. [34], in that it uses a more fine-tuned approach of stance detection, unlike the tool based sentiment analysis approach proposed in their work. Our work studies the ideological position of entities as well, alongside studying bias in media outlets. In this direction of analysing bias in the Indian mass media, we answer the following research questions in this section: (a) Which entities are most supportive or critical of the policies in the mass media? and (b) What is the ideological slant of media outlets regarding the economic and technology policies?

We have a significantly large dataset, and the models are trained only on an annotated subset of this dataset. The results and analysis in this section are presented after applying these trained models on the entire dataset.

8.1 Economic Policies

For this part, we consider our entire dataset of 5990 economic policy-related statements, out of which our model is trained only on 3855 annotated statements concerning four economic policies (350 Aadhaar, 1063 Demonetisation, 650 GST and 1792 Farmers' Protests). We filter out the neutral statements during the Step-1 (Stance Detection) of our classifier.

8.1.1 Ideological Position of Entities:

We use our ideology detection model to get the count of *Pro* and *Anti* statements from every entity. These counts provide us the overall stance or position of an entity towards a policy. From this analysis, we find that most leaders of the currently ruling Bharatiya Janta Party (BJP) (like *Narendra Modi*, *Arun Jaitley*, etc.) are more pro-policy (figure 8.1a) than the opposition leaders (like *Rahul Gandhi* (INC), *Mamata Banerjee* (TMC), etc.), who are more critical of the economic policies (figure 8.1b). This is expected since most of these economic policies were formulated or implemented during the term of the ruling party (2014-19).

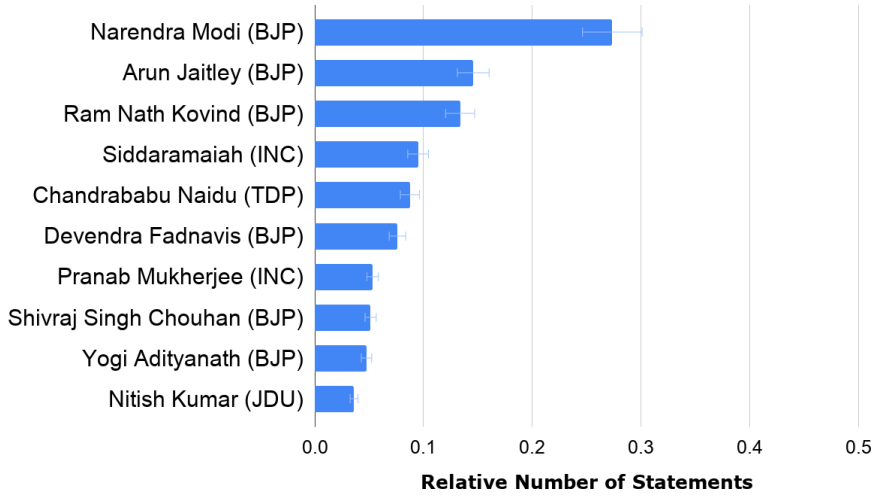
Our findings can also be corroborated by the statements that are made by these political leaders on the economic policies. For example, the prime minister *Narendra Modi*'s statement on Aadhaar enabled public distribution system (PDS),"*The government had detected 3.95-crore bogus ration cards, using technology and Aadhaar numbers to plug leakage in its social welfare programmes.*" shows his admiration towards the Aadhaar policy in detection of fake ration cards. On the other hand, the leader of opposition *Rahul Gandhi*'s statement on Aadhaar, "*For Congress, Aadhaar was an instrument of empowerment. For the BJP, Aadhaar is a tool of oppression and surveillance.*" is indicative of the opposition's criticism of the implementation of the policy by the ruling party (BJP). Similarly, we find the minister of finance *Arun Jaitley* making the statement, "*This is the positive impact of Demonetisation. More formalisation of economy, more money in the system, higher tax revenue, higher expenditure, higher growth after the first two quarters.*" in favour of Demonetisation. The leader of opposition *Mamata Banerjee* on the other hand made the statement, "*Demonetisation was not to combat black money. It was only to convert black money into white money for vested interests of the political party in power.*" indicative of her opposition towards the policy.

8.1.2 Ideological Slant of Mass Media:

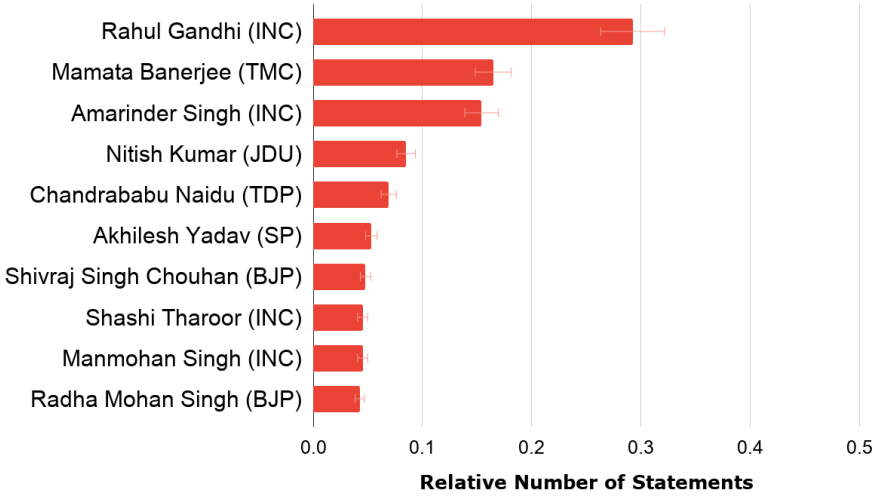
Figure 8.2 shows the predicted distribution of statements in various mass media sources. We see that for each media source, the number of pro-statements far exceeds the number of anti-statements. We consider two independent samples as the count of pro and anti statements classified for each of the six media sources. We conduct an independent t-test with 1-tailed hypothesis to analyse the variability in the pro ($M = 531.5$, $SD = 206.58$) and anti ($M = 202$, $SD = 76.14$) samples. The test indicates ($t(10) = 2.22$, $p = .025$) that the media sources under consideration demonstrated statistically significant ($p < .05$) pro-policy coverage than anti-policy coverage.

8.2 Technology Policies

For technology-related statements, the analysis is done on 2252 statements (517 Aadhaar, 653 E-Governance, 691 Digital India, and 391 Cashless Payments), while we only have the annotations of 812 statements. We also examine the predictions qualitatively on the unseen dataset. A small subset of the predictions has been shown in table 8.1. For more detailed results, please refer to the supplementary material [4].



(a) Pro Entities for Economic Policies



(b) Anti Entities for Economic Policies

Figure 8.1: Top 10 Entities (in terms of Pro/Anti statements) for both Economic Domain; Political Affiliation is denoted by (.)

8.2.1 Ideological Position of Entities:

A similar trend can be seen with the technology policies as well, where the leaders of the ruling party are found to favor technology strongly (figure 8.3a). For example, statements like these made by *Narendra Modi*, the Prime Minister of India, clearly indicate support towards a technology deterministic viewpoint: *"When poor farmers of villages have started adopting digital payments, now they (middlemen) have started spreading rumors."*. It is also noted that politicians of the ruling party are strong supporters of the technology policies rolled out by them. This statement made by *Ravi Shankar Prasad* on the Digital India policy clearly indicates the same: *"After coming to power, Prime Minister Narendra Modi gave the vision of Digital India as an important programme to transform India through the power of technology and bridge the digital divide."*. In contrast, leaders from the opposition parties (like *Rahul Gandhi* (INC), *Chandrababu Naidu* (TDP), etc.) are skeptical (figure 8.3b) of

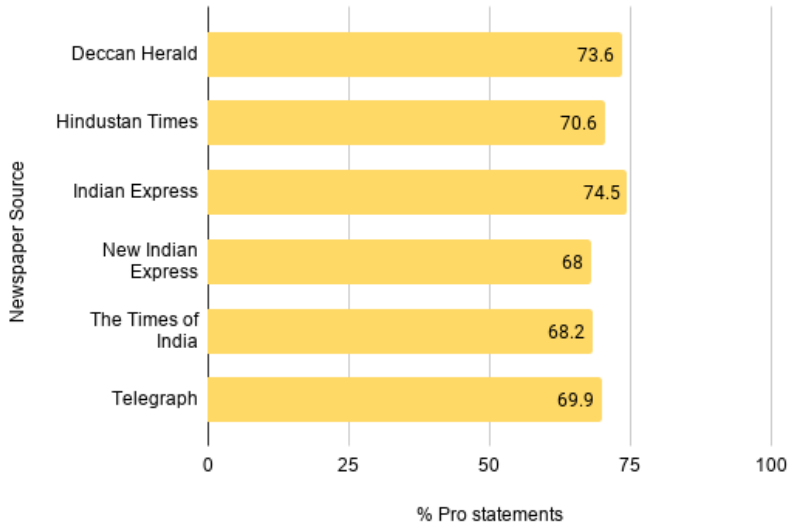


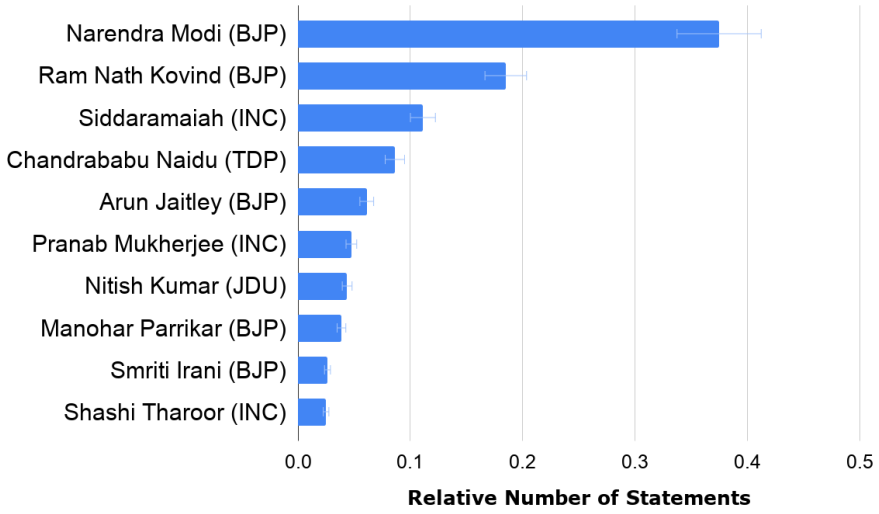
Figure 8.2: Percentage of Pro-Statements (economic) amongst various media sources

them. For example, the leader of the opposition, *Rahul Gandhi*, showed his skepticism towards the Digital India policy by making the statement: "*Digital India cannot become a euphemism for an Internet controlled by large remote corporations.*". Moreover, we see that the relative percentage of pro-technology statements is higher than the statements favoring the economic policies, which shows higher technology favoritism among policymakers.

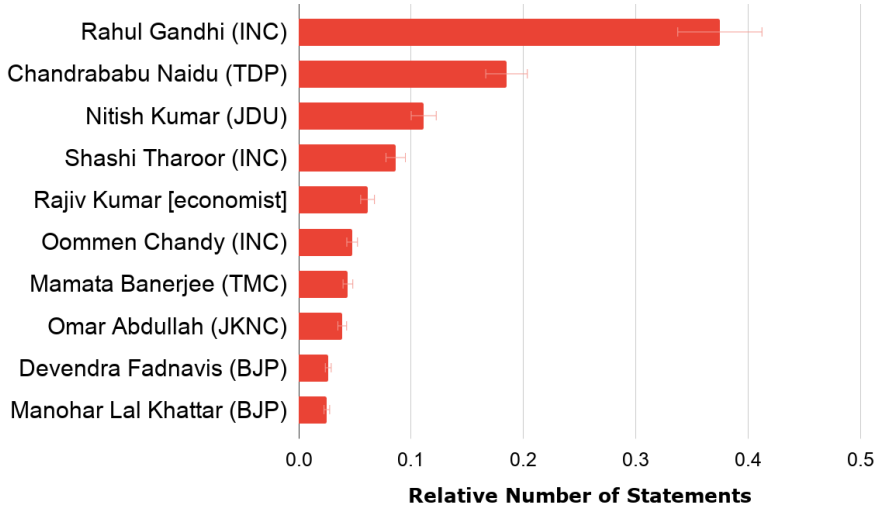
This tendency of policymakers to propose technology as a solution to many problems has been studied in previous works as well. Pal et al. [31] show how in order to justify the policy move of Demonetisation, the Prime Minister increasingly emphasized on the usage of digital cash and payment wallets by invoking patriotism, technical advancement, and projecting cashless payment as a one-shot solution to the problems of people. In another study, Pal et al. [30] discuss how the current Prime Minister of India branded his image as a tech-savvy modernizer on social media and other platforms.

8.2.2 Ideological Slant of Mass Media:

Similar to economic policies, the media mostly covers statements that reflect technology-determinism (figure 8.4). We conduct a similar t-test for pro-tech ($M = 249$, $SD = 81.14$) and anti-tech ($M = 63.67$, $SD = 14.76$) statements of these six media sources. The test indicates ($t(10) = 3.57$, $p = .025$) that the media sources under consideration demonstrated statistically significant ($p < .05$) coverage of tech-deterministic viewpoints as compared to the other side of the discourse. Additionally, these media sources seem to cover technology policies even more strongly than the economic policies (table 8.2). For each technology policy, the relative number of Pro, Anti and Neutral statements as covered by media are shown in Figure 8.5. It can be seen that as compared to other technology policies, more



(a) Pro Entities for Tech Policies



(b) Anti Entities for Tech Policies

Figure 8.3: Top 10 Entities (in terms of Pro/Anti statements) for Tech Domain; Political Affiliation is denoted by (.)

people are found to be neutral regarding the Cashless Payment technologies. Once again, most of our findings can be corroborated by earlier studies where the authors show how the Indian mass media is more keen on covering technology driven high-modernistic statements. For instance, Sen et al. [35] show how pro-technology aspects like *Development of Smart Cities* pertaining to the *Digital India* policy get a high coverage on mass media, while aspects that discuss the problems of the policy move get negligible coverage.

8.3 Generalizability

All the steps in our methodology until the model building stage are generic enough to be applied to any policy. The generalizability of our technology classifier model can be realized qualitatively from its performance on the 1440 unseen statements (table 8.1, more detailed

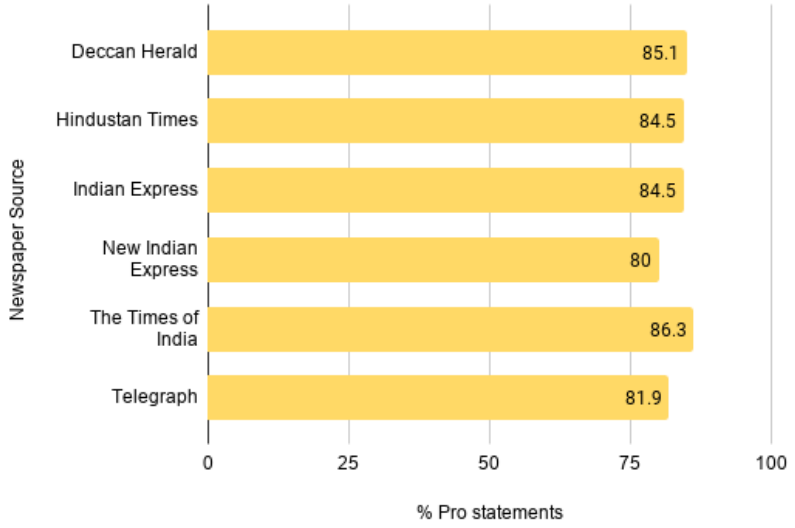


Figure 8.4: Percentage of Pro-Statements (technology) amongst various media sources

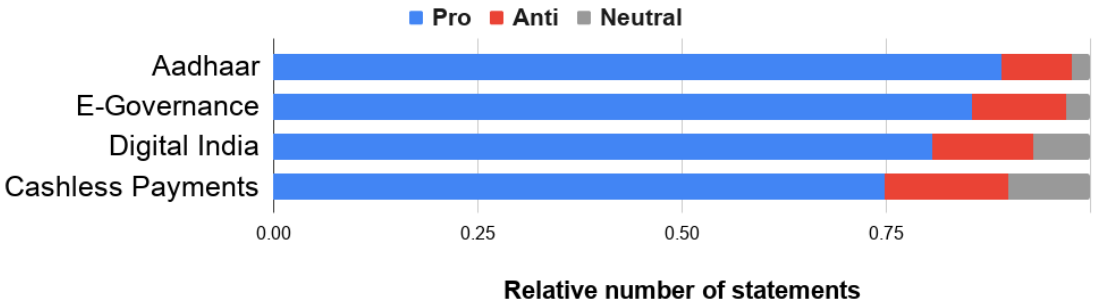


Figure 8.5: Normalised distribution of the statements amongst technology policies as covered by media

in the supplementary [4]). To investigate the generalizability of the economic classifier, we trained our model on three policies and tested on a different policy. We achieve a test performance (on 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% (F1-0.68) for Farmers' Protests after training on the remaining three policies. Our model performs reasonably well in this set-up except for Farmers' Protests, which may be because it requires a significantly different domain knowledge from the other policies. A misclassified statement like *"By 2015, the farmers of Devanahalli, Kolar and Chikkaballapur will have to migrate elsewhere, says Shivanapura Ram, a farmer from Devanahalli."* needs the context of farmers' migration, which results from poverty and unemployment. Our experiments thus show that our classifiers generalize well on unseen data, unless the domain of a new policy is significantly different than others. We are currently trying to improve our approach's generalizability further.

8.4 Misclassifications

We also analyzed the statements which have been wrongly classified. Some of them are as following :

8.4.1 Actually Anti but classified as Pro

1. Given the history of abuse by governments, it is right to ask questions about surveillance, particularly as technology is reshaping every aspect of our lives.
2. Though the country is spending huge amounts to import modern weapons from other countries, what we get most of the times are outdated technology systems.
3. Describing the Islamic State as one of the best users of Internet technology", has asked the armed forces to be prepared for future cyberwars while equipping soldiers for the physical battlefield.

8.4.2 Actually Pro but classified as Anti

1. Science is universal, but technology has to be local.
2. The commissioning of the facility also symbolizes the country's capability in establishing such world-class facilities wherein technology from outside is restricted or not available.

So, we observe that most of the misclassified statements are either themselves ambiguous (i.e. it's a bit confusing to know their gold label otherwise as well), or they have complicated sentence structures where one phrase is anti while the other is pro or they have some words which are generally representative of some sentiment e.g. restricted is generally a negative word and therefore the last sentence is being classified as Anti.

8.5 Sentiment Analysis with ID-RNN

As we know, sentiment analysis tools like Sentistrength work essentially like a bag of words model where they try to basically count on the presence and absence of certain sentiment-bearing words. We show in Table 8.3 that our deep learning classifier is able to outperform Sentistrength on several occasions.

We believe that our model can be further extended to determine the pro/anti stance of entire articles, in future. It also points towards the possibility of building a pro/anti government classifier, since the pro-policy statements generally show a high correlation with the pro-government ideology as the ruling members generally make them.

Table 8.1: Qualitative Analysis on the unseen technology-related dataset

Tech Policy	Pro	Anti
Aadhaar	*Observing that at present, over 113 crore residents in India have Aadhaar, Prasad said that Aadhaar is safe, let me say proudly that the data is secure.* *I want to really emphasise that Aadhaar platform is the biggest anti-corruption platform in the world.*	*The software sometimes fails to read fingerprints and Aadhaar details of beneficiaries, forcing them to return without their monthly quota of subsidised foodgrain, Modi said.* *They are not only opposed to EVMs, they have problems with technology, digital transactions, Aadhaar, GST, BHIM app.*
E-Gov.	*As many as 73,000 villages will be brought into banking network with the help of technology, Pranab Mukherjee said, adding actions are being taken.* *Our government wants to use technology to curb dishonesty and bring transparency in governance.*	*In a series of tweets, Vijay Mallya said the PM advocates about the use of technology while the enforcement agencies don't take his words seriously and refuse to use technology.* *Modi said that the example of breaking, addition, and twisting of technology is being seen in the form of social media.*
Digital India	*In a fresh push towards digital payments, Modi on Thursday told businesses to shun cash and go digital to bring transparency and root out black money.* *Modi said that in this age, more than physical connectivity there is need for information highways.*	*Gandhi said not a single person has benefitted by the prime ministers promises of controlling inflation, depositing Rs 15 lakh in the accounts of each citizen and upgrading certain towns to smart city.* *Retired town planner Ram said: the challenge of achieving the target of getting selected for smart city category is colossal.*
Cashless Payment	*Advocating for cashless transactions, Modi has said that the large volumes of liquid cash are a big source of corruption and black money.* *This will improve the functioning of toll plazas, digital payments, Das said.*	*Quelling fears over security of digital transactions, Babu said that the SBI had left no stone unturned to ensure that the data of the users were not compromised.* *It is the same as encouraging cashless initiatives without creating infrastructure post-demonetisation and recalling old notes without calibrating ATMs, said Jamshedpur Petroleum Dealers' Association General Secretary.*

Table 8.2: Relative Pro/Anti Predicted Distribution in Media Sources

Newspaper Source	**Economic**		**Technology**	
	Pro (%)	**Anti (%)**	**Pro (%)**	**Anti (%)**
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 8.3: Deep learning classifier v/s SentiStrength

Entity Statement	**Classifier Stance**	**Senti Strength Score**	**Explanation for SentiStrength Failure**
In order to help children suffering from an e-learning tool has also been developed by the National Institute of Mentally Handicapped and all the scholarships will be under one	Pro	-3	Mentions word 'Autism'
Bangalore has enormous energy and ideas to solve problems.	Pro	-1	Mentions word 'problems'
At UIDAI, we are very strict on privacy issues.	Pro	-2	Mentions word 'strict'
Earlier farmers used to get insurance of Rs 50,000 on death or permanent disability under Raj Sahakar Personal Insurance Scheme, which now has been increased to Rs 10 lakh.	Pro	-2	Mentions words, 'death' and 'disability'
Prime Minister Narendra Modi 's grand MSP increase for farmers is like applying a band-aid to a massive hemorrhage.	Anti	1	Sarcasm
Just because it is possible to hack a network did not mean that technology must not be deployed.	Pro	-1	Negation
Aadhaar would turn into a boon for marginalized or vulnerable groups to get access to many services with the help of single identity number opined.	Pro	-1	Mentions words like weak & marginalised sections
In the long term, this landmark step will increase the size of the official economy and reduce the shadow economy	Pro	0	Has both the words 'increase' and 'reduce'
The Pradhan Mantri Jan-Dhan Yojana provides a platform for changing the economic condition of our people.	Pro	0	Mentions 'change' rather than direct positive words like 'better'

Chapter 9

Conclusion and Future work

In this paper, we propose a framework to study the ideological biases existing in the Indian mass media, in terms of the statements covered by them on key economic and technology policies. We use an RNN based model to classify the statements made by influential elites on mass media into three classes of pro-policy, anti-policy, and neutral. Based on the coverage provided to these statements, we measure the ideological bias of different news-sources. Our findings indicate that the Indian news-sources generally cover the statements favoring the policies much more than those criticizing them, and take a pro-technology standpoint on technology policies. Our framework is generic enough to be applied to any other domain of ideology classification and presents a fine-tuned approach to detect the ideological position from policy discourse accurately. We believe that our framework and findings can serve towards pushing the Indian mass media towards greater self-regulation, enabling diversity in content publication, and educating the public about different viewpoints on key policies.

There are definitely some interesting directions to be pursued following the results we have achieved.

9.1 Tree LSTMs

Recursive Neural Networks sometimes fail to detect the change in stance in the tree structure. To facilitate this, TreeLSTMs [40] should be tried to better learn when to ignore the information of the sub-tree. TreeLSTMs are another kind of deep neural network that incorporates the hierarchical nature of sentences. TreeLSTM is a variant of ReNN where each node unit is an LSTM cell. There is a forget gate for each child of a node. Some papers claim to achieve better results using TreeLSTMs than those found by using recursive neural networks. Applying TreeLSTMs to our domain is definitely one interesting idea to work upon.

9.2 Fairness & Generalisability

We need to make better sense about the fairness of our model in terms of its predictions when assessed across different mass media sources. We need to find better ways to improve the generalisability of the classifier to new policies, and also need to design metrics that can quantitatively estimate the same.

9.3 Balanced Binary Trees

Also, it would be worth trying changes in the structure of the trees while incorporating Balanced Binary Trees [36] instead of the syntax trees. Balanced Binary Trees are constructed by building a balanced binary tree while using the words as leaves. They can make the trees more shallow so that the model can learn quickly and better. As mentioned in (Shi et al., 2018)[36] these trees give almost the same and even slightly better than syntax trees on some benchmarks.

9.4 Phrase-Level Labelings

The Recursive neural network model can be trained while incorporating the phrase level annotations. Although developing such a dataset is a challenge but (Iyyer et al., 2014)[14] show that phrase-level annotations improve the performance of the model. This is expected because in this case, the model is able to learn bottom-up from the phrase-level stance labels. Currently, in our work, we use the main root label and propagate it to all the phrases but this can potentially affect the accuracy of the classifier a lot since phrase labeling may not translate to the statement labeling in some cases.

9.5 Larger Dataset

The deep learning model performs quite well despite the limitation of the size of the dataset. However, it is certainly a promising idea to train the models over a much larger dataset as that would be able to truly unleash the powers of deep neural networks. It might require crowd-sourcing resources for being able to develop the ground truth for such a large dataset.

9.5.1 COVID-19 Relief Economy Policy

The government and the Reserve Bank of India have come out with a fiscal stimulus and a number of relief measures to protect the economy from the adverse impact of the ongoing Covid-19 crisis. Now these entail a large number of announcements that will directly or indirectly benefit the common people. This includes various economic relief measures target towards Micro, Medium and Small Enterprises ("MSMEs"), Defense Sector, Power Sector, Tax Measures, Ease of doing business measures, etc.

9.5.2 Aarogya Setu Tech Policy

Many countries have been attempting to supplement the work being done by health officials in tackling the COVID-19 pandemic with technology interventions. In India too, the Aarogya Setu app to track and alert those who physically come close to those who test positive for COVID-19, has aggressively been promoted by the government. Even as the govt pushes for aggressive adoption of its contact-tracing app, Aarogya Setu, privacy-focused groups such as the Internet Freedom Foundation (IFF) are raising alarm over its compliance with the globally-held privacy standards, while also recommending privacy prescriptions for these technology-based interventions.

9.6 Unstructured Datasets

It sounds very interesting to extend the idea to **unstructured datasets**. These days, social media has become a prominent platform for all kinds of discussions including the political & economic discussions. However, there are differences in the type of statements found in mass media to the ones found on Twitter. These differences have to be adequately addressed in the approach and consequently be learned by the model.

9.7 Sub-classification of biases

We can further sub-classify biases with respect to political parties:

- Newspaper speaking for or against a candidate
- Newspaper speaking for or against a party (generally found in opinion pieces): Here, the subject is implicit (the author writing an article where a party is mentioned)
- One party's candidate speaking against another party's candidate, or speaking for his/her party's candidate
- One party's candidate speaking against another party, or speaking for his/her own party
- One party's candidate speaking against another party's policy, or speaking for his/her own party's policy
- Newspaper speaking for or against a party's policy

9.8 Other Tweaks

Along with the above one could also explore some little tweaks in Iyyer et al. (2014)[14] such as using tree representation which is the representation of root concatenated with the average of the rest of the nodes in the tree instead of just root representation. Also, Iyyer et al. (2014)[14] report better accuracy by initialization of weight matrix to I/2, i.e, giving equal weight to each child of nodes initially. Moreover, it makes sense to try feature engineering. Some features like Part of speech tags are worth including in the list of experiments to be pursued.

Appendix A

SUPPLEMENTARY

A.1 Examples of different classes of By-Statements

Table A.1: Examples of Pro *by-statements* extracted

Policy	**By-Statements**
Aadhaar	*1. This makes it not only one of the most accurate, but soon to be the largest biometric system in the world," UIDAI Chairman Nandan Nilekani said in a statement here.* *2. Dalwai said the second phase of Aadhaar enrolments would be more stringent and foolproof, both at the enrolment level and data uploading level.*
Demonetisation	*1. Now there will be fewer cash transactions and an increase in digital currency, Jaitley said.* *2. In the wake of new currency note of a higher denomination of Rs 2,000 being introduced, such a cap on cash transactions would ensure that it does not become easy for hoarders to stack illicit savings in the higher denomination notes.*
GST	*1. The GST is beneficial for the poor people of the states represented by them because those states will economically benefit the most from GST, Modi said.* *2. Terming GST as the most historic reform in India, Modi said it will be implemented from next month in an apparent reference to the scheduled date of July 1.*
Farmers' Protests	*1. A sincere attempt will be made to provide water to standing crops of our farmers, Siddaramaiah said.* *2. Fadnavis told the villagers, we are making effort to ensure the investment cost in the agriculture is reduced to minimise the financial losses incurred by the farmers.*

Technology	*1. PM Modi also asked students to embrace the challenge of accepting newer technologies and said that they must not lose morale after facing setbacks while innovating new things.* *2. UIDAI Chairman Nandan Nilekani said the government is using state-of-the-art technology to ensure Aadhaar provides the best service for the people.*

Table A.2: Examples of Anti *by-statements* extracted

Policy	**By-Statements**
Aadhaar	*1. Modi said the banks should redouble efforts in financial literacy and seeding of Aadhar numbers with bank accounts needs to improve.* *2. Dalwai said the response received by UIDAI was too overwhelming and that there was an immediate need to enhance infrastructure, especially pertaining to processing of enrolments.*
Demonetisation	*1. It is common sense that if 86% of the cash is taken out of the market, there will be problems, said, adding that the government should have prepared itself better before making the announcement.* *2. Terming such acts as blatant fraud and cheating on the people, Amarinder once again demanded a review of the demonetisation policy to minimize the woes of the people".*
GST	*1. Addressing the traders, Gandhi said, Modi has enforced 'Gabbar Singh Tax' (GST), which has 'destroyed' your business.* *2. Gandhi also said if his party comes to power, it would change the GST and abolish the 28 per cent gst slab.*
Farmers' Protests	*1. Puttanniah said chief minister Siddaramaiah had said that the production of foodgrains in the state has declined by 15,000 tonnes, but he has not bothered to safeguard the farmers' interest.* *2. Gandhi had earlier said he will not let prime minister Narendra Modi sleep until all farmer loans have been waived off.*
Technology	*1. Fast-changing technologies like digitalization and climate change pose a challenge to the current as well as future generations, Modi said.* *2. "Matching such state-of-the-art systems could be a technological nightmare for Indian counterparts," Mohan said.*

Table A.3: Examples of Neutral *by-statements* extracted

Policy	**By-Statements**
Aadhaar	*1. Those without Aadhaar numbers will be given an opportunity to enrol for a UID number, Gautham said.* *2. If you want to use Aadhaar authentication, then your devices will have to be registered with us, Ajay Bhushan Pandey, CEO of UIDAI told PTI.*
Demonetisation	*1. In order to ease the cash situation, the government has formed a team of seven joint secretaries to monitor on a regular basis the shortage of the currency.* *2. Attorney General Mukul Rohatgi said that old currency notes worth Rs 8000 crore collected by DCCBs across the country have been allowed to be deposited in the Reserve Bank of India (RBI).*
GST	*1. Congress leader Gandhi had also tweeted government says 99 percent of goods will be at 18 percent GST.* *2. Mr. Ramakrishna Naidu, in the letter, said there was a reduction in effective rates of tax on several goods under the GST.*
Farmers' Protests	*1. Beyond that as the central government, I have nothing to say, Arun Jaitley told reporters when asked if the center will intervene on the issue of farm loan waivers and related protests.* *2. Confirming the date of the meeting, Chief Secretary Nagar also said that other proposals besides the loan waiver will be finalised on monday.*
Technology	*1. We plan to more than double the software robots to over 500 by end of this fiscal, Chanda Kochhar, MD, and CEO, ICICI Bank, said.* *2. Siddaramaiah said it is also proposed to establish artificial intelligence and robotics center at IIIT, Bengaluru at a cost of Rs five crore.*

Table A.4: Examples of Balanced *by-statements* extracted

Policy	**By-Statements**
Aadhaar	*1. As parties like BJD opposed the move to make Aadhaar mandatory for filing of income tax returns and making the PAN application, Jaitley said linking of Aadhaar with PAN was necessary as people have multiple pan cards and are using it as a tool for tax evasion.*

Demonetisation	*1. Hitting back at singh, who had called Demonetisation an organised loot and legalised plunder, Modi said, he (Manmohan) had perfected the art of bathing under a shower with raincoat on and so there was no blot on him despite all the scams that occurred during his tenure.* *2. The contraction in industrial production in December as per the latest data is the fallout of Demonetisation and expansion is expected in the coming months, finance minister Arun Jaitley said on Friday.*
GST	*1. Finance minister Arun Jaitley could announce tax incentives for individuals and firms to boost consumer demand, amid lingering uncertainty over the implementation of a nationwide goods and services tax, said one of the officials.* *2. Claiming victory of the Congress in its opposition of the land acquisition amendment bill moved by the center as six out of total nine important changes were taken back, Ram said the congress was in favor of the passage of goods and services tax (GST) bill but wanted four changes in its present form.*
Farmers' Protests	*1. At a review meeting held this week with district collectors to assess the drought situation, Chief Minister Devendra Fadnavis told the local administration, while the efforts that have helped reduce farmer suicides are laudable, we still have a formidable task ahead.* *2. Mr Modi said that his government stand by farmers and jawans of the country and blamed the Congress government for always befooling them and never thinking about their welfare.*
Technology	*1. Taking on the opposition and accusing it of misguiding people over cashless transactions, Modi said that while on one hand, they claimed that Rajiv brought about the telecom revolution, on the other they contradict themselves by saying that their countrymen don't have means for cashless transactions.* *2. In his address, Pranab Mukherjee said that demonetization has resulted in the temporary slowdown of the economy but as more and more transactions become cashless, it will improve the transparency of the economy.*

A.2 Coding Schema

Please refer to figures A.1 - A.4 and figure A.5 for the coding schema of economic and technology policies respectively.

Constituency	Statement primarily targets	Keywords	Examples	Normative Definition
Pro	providing new facilities/ linking or integration with other schemes (like delivery of subsidies)	money transfer, subsidies, direct benefit transfers, remittance facility, skype lite, integration with skype, delivery of subsidies, digital life certificate, provident fund, lpg subsidies, monthly pension	in a facebook post titled 'benefits of the aadhaar - where it stands today', arunjaitley said its use in the delivery of subsidies has helped saved rs 90,000 crore in the last few years till march 2018 by eliminating several duplicates, non-existent and fake beneficiaries. ssp lucknow deepakkumar said the gang used to advertise on social networking sites , inviting application from unemployed people willing to open customer service centre (csc) for issuance of aadhaar card, pan card, income certificate, caste certificate, domicile certificate and providing money transfer facility.	makes it easy to implement other schemes/policies like subsidiaries, pensions, opening of bank accounts; uses technology for identification of individuals, has led to technological developments; keeps a database of all the people living in India - be it citizens or not
	any statement citing advantages of aadhar/gpe/identification number/biometric data	jan dhan yojana, digital locker, unique identification document, primary identity,	the use of a fake aadhaar id will prove more hazardous to criminals as compared to spurious ration or election cards as biometric data can identify an individual with more certainty.	
	increased awareness in terms of other facilities like PAN, bank accounts, as a consequence of aadhar	opening bank accounts, voter card linking, Securities and Exchange board	eighty per cent of the two million individuals, who have been enrolled for the unique aadhar number so far, have requested for bank accounts, said nilekani while addressing the nasscom leadership forum 2011.	
	technological advancements due to/ related to aadhaar	modern security features, biometric recognition, BSID, Biometric Seafarer Identity Document	new delhi: microsoft ceo satyanadella on wednesday announced the integration of skype and aadhaar to launch the skype lite for low speed mobile internet connections, which will be supported by aadhaar and will enable technology that can empower people and organisations.	
	ease of use/ access/ registration/ time taken for implementation of the policy		we plan to increase the language versions in the future, so that residents who have access to the internet are able to view important information in a language that they are comfortable with, said person, chairman uidai while launching the multilingual website from uidai tech centre, bengaluru.	
Anti	new facilities being provided not reaching everyone/ having negative impact on someone/ rural poor getting excluded from PDS (public distribution system)	aadhar enabled starvations	if there are no banks or post office, how can one avail direct benefit transfer as there are lots of sc st people who dont have aadhar card, mamatabanerjee said.	benefits, as promised by linking of aadhar to different policies/schemes, not being equally transferred to everyone; difficulties in obtaining aadhaar and the issue of fake aadhar (security issues)
	improper/non-homogeneous execution of the policy		Making Aadhaar mandatory for availing subsidies was resulting in exclusion of beneficiaries and the government can not claim to have saved huge subsidies	
	privacy/security concerns related to aadhar/ leaking of information/ breach of trust	synthetic fingerprint generator, cloning of fingerprints, personal information privacy, Central Identities Data Repository, CIDR, hacking of UIDAI servers,	The report states Based on the numbers available on the websites looked at the estimated number of Aadhaar numbers leaked through these portals could be around million crore and the number of bank accounts numbers leaked at around million crore from the specific portals we looked at the upa legislation, arunjaitley said, was inadequate as it did not contain adequate safeguards on privacy and did not mention for which purpose the uid would be used. the report states: based on the numbers available on the websites looked at, the estimated number of aadhaar numbers leaked through these 4 portals could be around 130-135 million (13- 13.5 crore) and the number of bank accounts numbers leaked at around 100 million (10 crore) from the specific portals we looked at.	
	any activity related to the policy leading to violation of any law/ section		these websites and companies extending these unauthorised services, tantamount to violation under information technology act 2000, section 38 of aadhaar act 2016, and section 409 (criminal breach of trust) and section 420 (cheating) of ipc, according to uidai.	
	using the policy for political gains/ wrong information conveyed about the policy	money bill, right to privacy	The introduction of Aadhaar bill as a money bill was nothing but a brazen and mala fide attempt to bypass the approval of Rajya Sabha which holds an important place in the Constitutional and democratic framework of lawmaking	
	complications in the process of getting aadhar		While the enrolling agencies are paid up to Rs for a successful enrollment banks are charging about Rs for the same	
	complaints regarding aadhar/ requests for changes or amendment		fielding criticism from opposition parties such as the congress, the bjd and the tmc over aadhaar being made mandatory for income tax purposes, arunjaitley said in told the lok sabha, yes, we are making aadhaar mandatory. amendments must be made to the aadhaar act, 2016, after the enactment of a data protection law, in order to bring the act in consonance with the data protection framework, shashitharoor said.	
	unable to cover the whole population		the centres will facilitate enrolment of residents left out during the camps organised by registrars in the past and also facilitate biometric and demographic update, uidai chairman nandan nilekani said.	
	aadhar being misused/ used for illegal activities/ misuse in banking transactions			
	fake aadhaars/ credibility of aadhaar/petitions against aadhaar/ questions about the legitimacy or credibility of aadhaar		questioning the legitimacy of the uidai and the governments claim that aadhaar is the biggest tool to eradicate corruption in the country, justice (retd) k s puttaswamy, the main petitioner in the pil filed in the supreme court on uiadi, mathew thomas, managing trustee of the fifth estate, somasekhar v k, managing trustee of grahak shakti, and sunil abraham, ceo, centre for internet and society, on thursday announced that they would be soon filing a contempt petition in the apex court.	
	negative impact on national security			
	protests against the policy		16: chief minister mamatabanerjee today said trinamul mps would meet the prime minister soon to protest the centre's alleged efforts to make aadhaar cards mandatory by september for direct benefit-transfer schemes.	
Neutral	talks about the policy without carrying any ideology		rahul asked partymen to be involved in constructive work like helping to link people to the system and securing aadhar membership, union minister person and aicc general secretary person told reporters. the government was in favour of making Aadhar mandatory for all citizens saying why cant it be when per cent of the citizens already have it	

Figure A.1: Coding Schema - Aadhaar

Constituency	Statement primarily targets	Keywords	Examples	Normative Definition
Pro	measures to bring relief from difficulties faced due to demonetisation like recalibration of ATMs		for the urban population, das said recalibration of atms has already started and they will start dispensing the new rs 2,000 currency notes from today or tomorrow.	how demonetisation helped in curbing the issue of black money, fake currency and led to increase in digital payments and hence new technological developments
	helping to find frauds/ curb corruption/ black money	counterfeit banknotes, black money,	prime minister modi and the government had earlier claimed that demonetisation will put an end to black money, fake currency and terrorism.	
	stating advantages of demonetisation like lowering of interest rates, increase in public spending, benefits to the poor, economic growth		responding to noteban critics, arunjaitley said demonetisation has strengthen the economy and increased the government resources to fund poverty alleviation and infrastructure development programmes.	
	digital economy/ technological innovations	digital payments, cashless economy, mobile wallets,	after prime minister modi announced demonetisation on november 8, the current budget further aims to strengthen the countrys cashless economy.	
	equality in implementation			
	statements in support of the policy		economic affairs secretary subhashchandragarg said demonetisation and gst reflect long-term vision of the government and its ability to undertake massive structural reforms. repeating remarks he made in lok sabha a day earlier, modi said demonetisation should have happened in 1971 when indira gandhi was the prime minister.	
Anti	petitions against the policy/ request to change the policy/ people saying they are not support of the policy or want to rollback the policy	Akrosh Diwas	new delhi: west bengal chief minister mamatabanerjee and his delhi counterpart arvind kejriwal on thursday asked the government to scrap the demonetisation decision in three days or face a revolt. das said ever since the demonetisation announcement was made, a lot of representations have come to prime minister and finance minister person to ease withdrawal norms for wedding purposes.	Negative impact on economy and difficulties faced by the poor and middle class people due to note-ban; too many changes required for implementation like re-calibration of the ATMs all over the country
	statements saying that advantages claimed by the policy have not been delivered, eg, no fake currency recovered, does not help in curbing the problem of black money	information leak,	lalu said that modi may claim that it was his bold decision but in reality, demonetisation has taken the country back by many years.	
	difficulties faced by the people due to the policy, eg, standing in long queues, shortage of funds for various schemes, hurts small businesses	long queues, crisis-like situation, cash shortage, loss of jobs, decline in wages	atms, which reopened two days after prime minister modi announced demonetisation of two biggest currency notes, had people queued up since early morning.	
	problems arising/ failure to solve problems persisting as a consequence of demonetisation, for eg, not enough cash in ATMs		the finance ministry has urged people to withdraw only as much cash as they need from atms, and the ministry is looking into complaints about atms running dry, economic affairs secretary shaktikanta das said here on wednesday.	
Neutral	statements about the policy, without supporting or criticising it		GPE: Union Finance Minister PERSON on Thursday said Prime Minister PERSON has created a 'new normal' for the country's economy by breaking all the stereotypes of the past through his demonetisation move. bengaluru: demonetisation has been the talk of every home, street and by-lane of the country since prime minister modi announced the withdrawal of rs 500 and rs 1000 notes on november 8.	

Figure A.2: Coding Schema - Demonetisation

Constituency	Statement primarily targets	Keywords	Examples	Normative Definition
Pro	advantages of gst introduction like fighting tax evasion, corruption, black money, tax leakage		the long-delayed gst constitution bill was passed by parliament on august 8, marking a historic step for tax reforms which prime minister modi said was crucial for ending tax terrorism besides reducing corruption and black money while making consumer the king. in his 42-minute intervention, modi said the gst regime would compel traders to give proper bills to consumers and help fight corruption and black money.	Single tax for everything, less cumbersome, online system - technological development; the inclusion of small scale and unorganised sector leading to ease of doing business;
	regulation of unorganised sector under gst/ ease of doing business			
	proper implementation, ease of implementation eg filing returns		the technology used for filing gst returns was so simple that a student of class 10 or 12 could handle it."	
	statements giving evidence in support of the policy		speaking about the power of unity, modi said, in august 2016, parties having intense political rivalry, parties which do not let go any chance to attack the other, all came together to pass the gst bill. insisting that states have nothing to fear, arunjaitley said, the structure of the gst council is such that states have the veto power. simultaneously the gst council will work on the functional modalities for implementation such that same person is not assessed by both centre and states, arunjaitley said.	
	Addressal of grievances/ queries related to gst, by the govt		about gst implementation, shailesh patwari, president of the gcci, said, "the government has been addressing all our queries ever since gst was implemented.	
	only one tax instead of many taxes like before/ common national market			
	gst being an online system - online taxation database, makes refunding easy, reducing litigation		countries that rolled out gst have witnessed higher economic growth, says udit ram , adding that an online database of taxes will increase transparency. as refunding is a big issue, the gstn decided on complete online refunding from september 24 this year from a single source, either by the central gst or state gst, modi said.	
Anti	difficulties in being GST-compliant; complicated laws - difficult to understand		it is also highly expected that many sections of the industry may not be fully compliant with various requirements of gst law from day 1 itself and would require a transitioning period of few months, post the go-live date," said rajeev dimri, bmr and associates llp leader (indirect tax). therefore, excise duty collection this quarter may fall unless the government provides higher percentage of deemed credit (currently proposed at 40 per cent of cgst), jain said. addressing a seminar on compliance procedures under gst organised by the federation of andhra pradesh and telangana chambers of commerce and industry (ftapcci) here, mr. ramakrishnudu said gst had not yet been fixed for 10 commodities.	difficult being gst compliance, due to installation of new softwares, purchase of new equipment, complex gst laws; negative impact on small scaled businesses and hence the overall economy
	any statement stating that people/businesses/organisations have issues with gst and/or want to make changes to the gst laws		congress leader rahulgandhi had also tweeted government says 99 per cent of goods will be at 18 per cent gst. we have set up a gst help desk at the gujarat chamber of commerce and industries (gcci) and are representing the issues of specific industries to the government," said shailesh patwari, the gcci president. rajender said that while supporting the implementation of gst (goods and services tax), telangana urged the centre to exempt paddy, petrol and diesel from the new tax regime. i have already asked the state chief secretary to send a strong letter to the union finance ministry elaborating on our objections to accepting the gst in its present form, bengal chief minister mamatabanerjee said at a district-level meeting in her state. former finance minister and congress leader p chidambaram former finance minister and congress leader p chidambaramthe proposed multiple rate gst structure will be disastrous and nothing more than same old vat rates in a new shape, former finance minister p chidambaram said on monday.	
	state-wise non-uniform implementation		in a statement, its president n. jegatheesan and senior president s. rethinavelu said that the trade and industry was apprehensive that when the gst was introduced under single country, single market, single tax principle, the revenue neutral rate might be fixed by individual states at a higher scale.	
	everything being not covered under gst		that said, it remained on a weak growth trajectory amid reports that the goods and services tax (gst) was still hindering efforts to secure new clients," said aashna dodhia, economist at ihs markit, and the author of the report.	
	micro, small, medium, local enterprises will have higher tax burden/ would be adversely affected due to gst		in case the gst is not extended to the state, the businesses will be crippled as no trader from jammu and kashmir will be able to do business with their counterparts from other parts of the country," drabu said, adding the consumer in j-k will be the worst hit due to double taxation. gandhi said, apart from fabrics import, manufacturers of fabrics are comfortable with gst rates of 5 and 12 on mmf yarn and that any increase by the government in gst rates will lead to inflation of fabric prices and the whole value chain will be disrupted. accusing the government of bringing gst to favour the corporates, chidambaram said his party congress had to speak for the common people, who were the "third factor" besides the centre and states, that would be affected by taxes.	
	complexity in running bussiness under gst laws		expressing his opinion with regard to the complexities involved in running businesses under gst laws, chidambaram said, you cant have this fantastic nonsense of asking people to file three returns a month in every state, which means if im doing business all over india, we have to file thousand returns.	
	unemployment due to gst		in bhilwara, gandhi said demonetisation and gst left lakhs of people unemployed.	
	gst is an online system - increased costs due to software purchase for gst implementation; tough for smaller business to adapt to online system			
	difficulties related to (hurried) implementation in the middle of financial year			
Neutral	any statement which talks about the policy but neither in a positive nor negative manner		in gst, there is no centralised exemption but if a state wants, it can refund the state gst for promoting regional cinema, arunjaitley said.	
	any statement containing both pro and anti		the government and opposition parties were thursday locked in a bitter war of words with finance minister arun jaitley citing a spike in the number of taxpayers to staunchly defend demonetisation while congress president gandhi said the note ban cost 1.	

Figure A.3: Coding Schema - GST

Constituency	Statement primarily targets	Keywords	Examples	Normative Definition
Pro	good revenue generation good price for crops protect interests of poor farmers increased crop production	swaminathan commission report more price of crops kisan channel more liquidity introduction of msp increased Value-Added per Hectare increased productivity abolishing intermediaries through land reforms expanding institutional credit support to poor farmers etc.	1. the chief minister also demanded that the farmers be paid the due price for their produce and in toto implementation of the swaminathan commission report. 2. however, siddaramaiah earlier told representatives of various farmers organisations that the online system had been working effectively and helped lakhs of farmers get judicious price for their produce at apmcs. 3. making a strong outreach to the crisis-ridden farming community, modi said the kisan channel is aimed at helping drive farm income by supplying key information, such as weather, market and demand. 4. about the steps taken by government to protect farmers like setting up of enforcement cells at district level to keep a check on private money lenders and repayment of dues to sugarcane farmers, siddaramaiah said government is planning to constitute an expert committee	how government is introducing plans that serve to protect the interests of the farmers - either by giving a good price to crops, crop insurance, financial assistance, waiving loans, countering suicides, cooperation, etc.
	increased compensations to damages more incomes to farmers financial relief financial assistance	increased compensation increased budget in agriculture	1. union urban development minister m.venkaiahnaidu in gandhinagar on monday announced that the centre will decide on relaxing guidelines to pay compensation to farmers who sustained 25 per cent to 50 percent damage to their crops. 2. modi announces 50 pc increase in compensation for distressed farmersnew delhi, april 09, 2015, dhns: 2:28 istin a reprieve to farmers whose crops have been hit by recent unseasonal rain, prime minister narendramodi on wednesday announced a 50 per cent increase in compensation and also eased criterion for providing financial relief. 3. expressing concern of farmers' suicides and the plight of the poor people in the country, modi asked the bankers to be considerate in providing financial assistance to poor. 4. union home minister rajnathsingh on friday said the nda government was working hard to improve the condition of farmers, claiming it was the first government ever to have resolved to double farm incomes by 2022. 5. finance minister arun jaitley this month said the government will spend a record amount on rural areas and farming to help double farmers income by 2022. 6. chief minister yogiadityanath on saturday announced an ex gratia of rs 4 lakh to families of the deceased and directed district magistrates to assess crop damage and identify farmers eligible for financial aid. 7. pm modi said the budget for a 'new india' has a roadmap to transform the agriculture sector of the country.	
	increased cooperation		1. india and israel will strengthen the existing cooperation in agriculture, technology and security, modi said at a joint media event with netanyahu. 2. during the talks, modi said india and vietnam agreed to deepen trade and investment ties in sectors like oil and gas exploration, renewable energy, agriculture and textiles.	
	policies against crop loss avail crop insurance	pradhan mantri fasal bima yojana accident insurance coverage	1. crop insurancepatna: union agriculture minister radhamohansingh on friday said the centre has extended to august 31 the last date to enrol in the pradhan mantri fasal bima yojana for bihar, aimed at enabling farmers to avail insurance cover against crop loss because of natural calamities. 2. prime minister modi had announced the pm crop insurance scheme, the pradhan mantri fasal bima yojana, in 2015. 3. the pmfby has expanded the insurance cover to crop loss due to hailstorm, land slide and inundation of farms, modi said adding the scheme would also extend to post-harvest losses due to cyclonic and unseasonal rains.	
	providing loan waivers to farmers	farm loans loan waivers or assistance debts	1. prime minister modi too had said in his campaign speeches that writing off farm loans would be among the foremost tasks of the bjp government in uttar pradesh. 2. let centre also waive farm loans, cm sayshubballi, oct 02, 2015, dhns: 0:35 ista day after the state government announced waiver of interest on farmers medium- and long-term loans borrowed from co-operative banks, chief minister ram asked bjp leaders to pressure the union government to waive farmers loans borrowed from the nationalised banks, or, at least the interest. 3. gandhi said, if farmers are in distress, waiver is the immediate step that needs to be taken.	
	introducing new technology start-up campaign	technology in agriculture Modernizing agricultural sector improved agricultural inputs like HYV seeds, fertilizers etc. promote agricultural research and training facilities close linkage between research institutions and farmers	1. chief minister siddaramaiah announced cloud seeding, in the backdrop of five consecutive droughts. 2. referring to two young entrepreneurs working in sikkim on an agriculture start-up, pm modi said, it is a great way of linking agriculture to the start-up india campaign. 3. however, siddaramaiah earlier told representatives of various farmers organisations that the online system had been working effectively and helped lakhs of farmers get judicious price for their produce at apmcs	
	policies concering farmers suicides	relief funds, water management	1. farmers' suicide; cm siddaramaiah to lead all party delegation to pmhubballi(karna), aug 5, 2015, (pti) 16:53 istkarnataka chief minister siddaramaiah today said he will lead an all party delegation to the prime minister to seek more relief funds in the wake of farmer suicides, unseasonal rains and drought like situation in certain parts of the state. 2. chief minister devendra fadnavis on friday said an integrated scheme to ensure economic sustenance for small and marginal farmers was being worked out to put an end to farmer suicides in maharashtra. 3. (pti photo)calling for a relook at the proposal to link rivers in the country, prime minister narendramodi on saturday said that if water management is given importance, there would be no farmer suicides.	
	checking environmental degradation	stubble burning droughts	1. punjab chief minister amarindersingh said the project would help in addressing the problem of the environmental pollution due to stubble burning besides providing additional income to farmers by helping turn the unmanageable agro-waste into raw material for producing bio fuel. 2. the ruling bjp does seem to have realized that climate change and droughts are a growing concern, and prime minister narendramodi announced a separate ministry for water called jal shakti.	
Anti	failed to protect interests of farmers not considerate towards farmers ignoring farmers interest		1. siddaramaiah said: agriculture universities, which came into being to help develop the agriculture sector, have failed to protect the interests of farmers. 2. nationalist congress party leader ajit pawar said that the bjp-led govenrment was working in the interest of a handful of industrialists and would do nothing for the farmers.	how the government schemes inhibit/ignore to protect the interests of farmers with poor loan waiver schemes, hike in prices, poor financial assistance, no relief funds, dearth of resources, etc.
	poor price of crops low crop productions		1. congress chief rahulgandhi had said that giving farmers rs 17 a day was an insult to all that they stood for. 2. chief minister ram had said that the production of foodgrains in the state has declined by 15,000 tonnes, but he has not bothered to safeguard the interest of farmers.	
	loan waiver schemes		1. taking a dig at farm loan waiver schemes in rajasthan and mp where congress had won in recent state elections, pm modi said these agri schemes had fallen flat. 2. only 80,000 of four lakh farmers got crop loans this kharif season under the governments restructured crop loan scheme. 3. the congress government had failed to deliver on its promise to waive farm loans, sad president sukhbirsinghbadal said monday. 4. gandhi said the modi government has waived off loans of big industrialists but when farmers want the waiver of their loans, where the amount involved in much less, jaitley would not oblige. 5. finance minister arunjaitley (file) finance minister arunjaitley (file)soon after finance minister arunjaitley announced the centres disassociation with loan waivers for farmers, the congress party on tuesday lashed out at the former, saying the bharatiya janata party (bjp) must not make promises they cannot deliver.	
	price hike	surge in prices of fertilisers, diesel, fdi	1. this price hike will prove to be very bad for the farmers of the state where there is drought-like situation, modi said. 2. criticising the diesel price hike,chief minister narendramodi on thursday said it would affect farmers the most. 3. at a rally in east midnapore yesterday, mamatabanerjee told the gathering, largely constituting farmers, that fdi and fertiliser price hikes would ruin them and leave them landless.	
	farmers' strike		1. pticongress leaders on sunday termed union agriculture minister radha mohan singhs remark on the ongoing farmers strike, or gaon bandh, as insensitive and full of arrogance. 2. we have left with no other option except to go for agitation, said ram , one of the farmers.	
	lack of resources	shortage of water supply cattle	1. karnataka was not releasing water for its own farmers, as the available storage has been reserved to meet the drinking water needs of bengaluru and other cities, ram told the team of 20 farmers associations from tamil nadu, led by former rajya sabha mp and the leader of dmks farmers wing, kp ramalingam. 2. the centres new notification imposing restrictions on cattle trade will cause an enormous financial burden on farmers, thereby further contributing to the agrarian crisis, karnataka chief minister siddaramaiah has said in a letter to prime minister narendra modi.	
Neutral	any statement which talks about the policy but neither in a positive nor negative manner talks about the condition of the farmers without any take on policy		1. stating that he was himself a son of a farmer, siddaramaiah said his government would make farmer welfare its priority. 2. speaking at the opening ceremony for the ninth edition of the hugely-popular league, shukla said, the ipl and bcci are committed to the welfare of the farmers. 3. modi says that the five aspects of congress rule are dynasty politics, corruption , rampant lawlessness, agrarian distress and division of society. 4. in his address broadcast through air stations across the state, siddaramaiah said 70 farmers had committed suicide since june which was a matter of concern and the suddenrise had remained an enigma to the government itself.	
	any statement containing both pro and anti		1. but even as mamatabanerjee announced wednesday, from the mammoth stage set up for the first singur utsav that she was against land acquisition and will never allow forcible acquisition in bengal ever again, even as she vowed to protect farmers and their lands, in the same breath she said that her government will push for industrial investment in the state.	

Figure A.4: Coding Schema - Farmers' Protests

Constituency	Statement primarily targets	Keywords	Examples	Normative Definition
Pro	upgradation of existing methods/ improvement in existing systems using technology/Digital India/Make in India	make in india, digital india, bharat broadband network, e-governance, bharatnet, aadhaar, aadhaar enabled payment system, accessible india campaign mobile app, BHIM(bharat interface for money), centre for excellence for IoT, etc (check https://digitalindia.gov.in/di-initiatives for more) , smart cities, IITs/IIITs/NITs	(ap photo)prime minister modi on tuesday here asked the iits, iiits and other technical institutions to create an atmosphere for boosting up the make in india campaign so that the country could not only step in a big way to manufacture defence equipment, but also lead the world in ensuring cyber security and supply skilled manpower globally in the next few years.	Any statement which shows the belief that technology is the solution to anything and everything and only good can come out of it. "Technology is the new electricity"
	help any entity(organization/section of society/person) in a positive manner	make in india, digital india, bharat broadband network, e-governance, bharatnet, aadhaar, aadhaar enabled payment system, accessible india campaign mobile app, BHIM(bharat interface for money), centre for excellence for IoT, etc (check https://digitalindia.gov.in/di-initiatives for more) , smart cities, IITs/IIITs/NITs	interacting with beneficiaries of the various digital india efforts, modi said the initiative was launched with an objective of bringing benefits of technology to people, especially in rural areas.	
	help in innovation/ISRO missions		modi said the digital technology was introducing transparency and eliminating corruption through innovations such as the government e-marketor gem.	
	help in awareness/removing illiteracy/debunking myths, superstitions			
	leading to improved security; helpful for defense/military/stealth/aircrafts/missiles		Kalvaris induction in the Navy is a big step in defense preparedness, for undertaking the project to construct the six submarines with technology transfer from the Naval Group (Formerly DCNS) of France	
	help to create opportunities/jobs/business-expansion/employment		technology is defining competitiveness and power in the new world and it is creating boundless opportunities to transform lives, modi said as he began his two-day visit to the country.	
	innovations in different fields/new developments		use technology such as mobile apps, G-tagging and engage citizens to accelerate the implementation of works related to rural roads There are huge opportunities for partnership in every area we can think of agriculture, agro-processing, resources, energy, finance, infrastructure, education, and science and technology The success in agriculture needs to be extended to eastern India and scientific and technological interventions are required to make this a reality	
	ease of use/accessibility	easily accessible		
	fostering sustainable development		addressing a gathering after delivering a lecture on space technology and societal applications, a s kirankumar , chairman, isro, said on monday that isro is working on designing reuseable satellite launch vehicles in an effort to cut down cost.	
	ensuring safety and law and order/ detection of frauds		courts could use video conferencing to communicate with government officials instead of summoning them to appear in court, saving time and money.	
	resulting in more globalisation/ global connectedness/ enhancing ties/ increased collaboration with other countries		India and the UK can leverage technological prowess to create new opportunities while seeking the UK businesses to invest in defense, manufacturing and aerospace sectors in India.	
Anti	resulting in harm to any entity/organisation/society			Any statement showing the belief that using technology will only bring more harm than benefit
	loss in human values/ traditions/ intelligence		Technological progress without an equivalent progress in human institutions can doom us. adding that such technology "requires a moral revolution as well".	
	cyber-wars/cyber-attacks/security/privacy abuse/Cyberterrorism/surveillance/hacking/ data leakage		Global terrorism has evolved over time and terrorists are now using modern technology and devices while the national and international efforts to counter them have become outdated	
	using outdated tech/ issues with adaptation from region to region		Global terrorism has evolved over time and terrorists are now using modern technology and devices while the national and international efforts to counter them have become outdated	
	large budget/wastage of money/frequent failures/futile attempts		expressing serious concern over reports of failure of evms at several places across the state, babu said he had been saying right from the beginning that evms were unreliable as they were prone to technical glitches.	
	unequal distribution/unavailability			
	negative effects on lifestyle/health		at a colourful ceremony, which included a presentation of india's indigenous martial art forms, dances and sports, modi said technology has contributed to a sedentary lifestyle.	
	impact on environment		The society must concentrate to green their lifestyle (sic) and lessen the negative impact of technology on natural environment The Goa government will not allow more coal to be handled at Mormugao Port Trust until companies involved in its transportation come out with technology to control the pollution caused by the fossil fuels movement	
	difficult to integrate			
Neutral	talks about technology without carrying any ideology		the brightest and best in every corner of India should have the opportunity to excel in science, this will ensure that our youth get high-end training exposure to the best of science and technology to make them job-ready in a competitive world. Yoga is both an art as well as a science and has amazing curative and preventive powers for the well-being of humanity, on International Yoga Day on Sunday. Niti Aayog to examine global standards of technology for infrastructure creation and their feasibility in India.	

Figure A.5: Coding Schema - Technology

Appendix B

IMPORTANT CODE SNIPPETS

B.1 By-Statement Extraction and Entity-Specific Coverage Analysis

```
def entitySpecificCoverageAnalysis(doc_set, entity_keywords, entity_name, e_aliases):
    '''
    Finds the sentences that are about or by the entity
    :param doc_set: set of sentences
    :param entity_keywords: keywords as to which entity to identify in the sentence.
    :return: onTarget_sentences, byTarget_sentences, removed_sentences, onTargetTopic, byTargetTopic
    '''
    sNLP = StanfordNLP()
    onTargetArticles = []
    byTargetArticles = []
    removedArticles = []
    short_entity_name = ''.join(entity_name.split()).lower()
    entity_keywords.append(short_entity_name)
    for i in range(len(doc_set)):
        text = preprocesstext(doc_set[i])
        for alis in e_aliases:
            text = text.replace(' ' + alis.lower() + ' ', ' ' + short_entity_name + ' ')
            text = text.replace(' ' + alis.lower() + '. ', ' ' + short_entity_name + ' . ')
            text = text.replace(' ' + alis.lower() + ', ', ' ' + short_entity_name + ' , ')
        try:
            pos_text = sNLP.pos(text)
        except json.decoder.JSONDecodeError:
            print('JSON_Decode_Error: ', text)
            continue
        parse_text = sNLP.dependency_parse(text)
        state1 = False
        state2 = False
        for pt in parse_text:
            if ((pt[0] == 'nsubj') or (pt[0] == 'nmod') or (pt[0] == 'amod') or (pt[0] == 'dobj')) and ( (pos_text[pt[1] - 1][0] in entity_keywords) or (pos_text[pt[2] - 1][0] in entity_keywords)):
```

```
                if ((pt[0] == 'nsubj') and ( pos_text[pt[1] - 1][0] in fixed_keywords or pos_text[pt[2] - 1][0] in fixed_keywords)):
                    state2 = True
                else:
                    state1 = True
        if state1:
            onTargetArticles.append(text)
        if state2:
            byTargetArticles.append(text)
        else:
            removedArticles.append(text)
    return (onTargetArticles, byTargetArticles, removedArticles)
```

B.2 Fine-tuning Word2Vec

```
def get_domain_model(corpus, word2vec_model):
    # check size of embedding of word2vec
    embedding_dim = word2vec_model.vectors[0].shape[0]
    domain_model = gensim.models.Word2Vec(size=300, alpha=0.025, window=5, min_count=2, max_vocab_size=None, sample=0.001, workers=4, min_alpha=0.0001, sg=0, hs=0, negative=5, ns_exponent=0.75, cbow_mean=1)
    domain_model.build_vocab(corpus)
    total_examples = domain_model.corpus_count
    domain_model.build_vocab([list(word2vec_model.vocab.keys())], update=True)
    domain_model.intersect_word2vec_format(pretrained_embeddings_path, binary=True, lockf=lock_factor)
    domain_model.train(corpus, total_examples=total_examples, epochs=1)
    return domain_model
```

B.3 Recursive Neural Network

```
class RecursiveNN(nn.Module):
    def __init__(self, word_embeddings, vocab, embedSize=300, numClasses=2, beta = 0.3, use_weight = True, non_trainable = non_trainable):
        super(RecursiveNN, self).__init__()
        self.embedding = nn.Embedding.from_pretrained(word_embeddings)
        self.embedding.weight.requires_grad = True
        if non_trainable:
            self.embedding.weight.requires_grad = False
        else:
            self.embedding = nn.Embedding(len(vocab), embedSize)
        self.embedding = nn.Embedding(len(vocab), embedSize)
        self.W = nn.Linear(2*embedSize, embedSize, bias=True)
```

```
12          self.nonLinear = torch.tanh
13          self.projection = nn.Linear(embedSize, numClasses, bias=True)
14          self.nodeProbList = []
15          self.labelList = []
16          self.loss = Var(torch.FloatTensor([0]))
17          self.V = vocab
18          self.beta = beta
19          self.use_weight = use_weight
20          self.total_rep = None #
21          self.count_rep = 0 #
22          self.numClasses = numClasses
23
24      def traverse(self, node):
25          if node.isLeaf:
26              if node.getLeafWord() in self.V:  # check if right word is in
    vocabulary
27                  word = node.getLeafWord()
28              else:  # otherwise use the unknown token
29                  word = 'UNK'
30              currentNode = (self.embedding(Var(torch.LongTensor([int(self.V[
    word])]))))
31          else: currentNode = self.nonLinear(self.W(torch.cat((self.traverse(
    node.left),self.traverse(node.right)),1)))
32          currentNode = currentNode/(torch.norm(currentNode))
33          assert node.label!=None
34          self.nodeProbList.append(self.projection(currentNode))
35          self.labelList.append(torch.LongTensor([node.label]))
36          loss_weight = 1-self.beta if node.annotated else self.beta
37          self.loss += (loss_weight*F.cross_entropy(input=torch.cat([self.
    projection(currentNode)]),target=Var(torch.cat([torch.LongTensor([node.
    label])]))))
38          if not node.isRoot():
39              if self.total_rep is None:
40                  self.total_rep = currentNode.data.clone()
41              else:
42                  self.total_rep += currentNode.data.clone()
43              self.count_rep += 1
44          return currentNode
45
46      def traverse_test(self, node):
47          if node.isLeaf:
48              if node.getLeafWord() in self.V:  # check if right word is in
    vocabulary
49                  word = node.getLeafWord()
50              else:  # otherwise use the unknown token
51                  word = 'UNK'
52              currentNode = (self.embedding(Var(torch.LongTensor([int(self.V[
    word])]))))
```

```
53          else: currentNode = self.nonLinear(self.W(torch.cat((self.
    traverse_test(node.left),self.traverse_test(node.right)),1)))
54          currentNode = currentNode/(torch.norm(currentNode))
55  #           assert node.label!=None
56          self.nodeProbList.append(self.projection(currentNode))
57          loss_weight = 1-self.beta if node.annotated else self.beta
58
59          if not node.isRoot():
60              if self.total_rep is None:
61                  self.total_rep = currentNode.data.clone()
62              else:
63                  self.total_rep += currentNode.data.clone()
64              self.count_rep += 1
65          return currentNode
66
67      def forward(self, x):
68          self.nodeProbList = []
69          self.labelList = []
70          self.loss = Var(torch.FloatTensor([0]))
71          self.traverse(x)
72          self.labelList = Var(torch.cat(self.labelList))
73          return torch.cat(self.nodeProbList)
74
75      def getLoss(self, tree):
76          nodes = self.forward(tree)
77          predictions = nodes.max(dim=1)[1]
78          loss = self.loss
79          return predictions,loss
80
81      def getRep(self, tree):
82          self.count_rep = 0
83          self.total_rep = None
84          self.nodeProbList = []
85          self.labelList = []
86          self.loss = Var(torch.FloatTensor([0]))
87
88          root_rep = self.traverse(tree)
89
90          return (torch.cat((root_rep,self.total_rep/self.count_rep),1)).data.
    numpy().T.flatten()
91
92
93      def predict(self, trees):
94          pbar = progressbar.ProgressBar(widgets=widgets, maxval=len(trees)).
    start()
95          preds = []
96          for j, tree in enumerate(trees):
97              nodes = self.forward(tree.root)
```

```
98              predictions = nodes.max(dim=1)[1]
99              preds.append(predictions)
100             pbar.update(j)
101         pbar.finish()
102         return preds
103
104
105     def evaluate(self, trees):
106             pbar = progressbar.ProgressBar(widgets=widgets, maxval=len(trees)
    ).start()
107             n = nAll = correctRoot = correctAll = 0.0
108             tp = [1e-2]*self.numClasses
109             fp = [1e-2]*self.numClasses
110             fn = [1e-2]*self.numClasses
111             f1 = [0.]*self.numClasses
112             for j, tree in enumerate(trees):
113                 predictions,_ = self.getLoss(tree.root)
114                 correct = ((predictions.cpu().data).numpy()==(self.labelList.
    cpu().data).numpy())
115                 correctAll += correct.sum()
116                 nAll += np.shape(correct.squeeze())[0] if np.size(correct)!=1
     else 1
117                 correctRoot += correct.squeeze()[-1] if np.size(correct)!=1
    else correct[-1]
118                 for i in range(self.numClasses):
119                     size = np.size((predictions.cpu().data).numpy())
120                     if size!=1:
121                         pred = (predictions.cpu().data).numpy().squeeze()[-1]
122                         actual = (self.labelList.cpu().data).numpy().squeeze
    ()[-1]
123                     else:
124                         pred = (predictions.cpu().data).numpy()[-1]
125                         actual = (self.labelList.cpu().data).numpy()[-1]
126                     if pred==i and actual==i:
127                         tp[i]+=1
128                     elif pred==i and actual!=i:
129                         fn[i]+=1
130                     elif pred==i and actual!=i:
131                         fp[i]+=1
132                 n += 1
133                 pbar.update(j)
134             for i in range(self.numClasses):
135                 p =(1.0*tp[i]/(tp[i]+fp[i]))
136                 r =(1.0*tp[i]/(tp[i]+fn[i]))
137                 f1[i] = (2*p*r)/(p+r)
138             pbar.finish()
139             return correctRoot / n, correctAll/nAll, f1
140
```

```
141     def eval_sent_lvl(self,trees,clf):
142         pbar = progressbar.ProgressBar(widgets=widgets, maxval=len(trees)).
    start()
143         n = nAll = correctRoot = correctAll = 0.0
144         X_predict = []
145         Y_gold = []
146         for j, tree in enumerate(trees):
147             tree_rep = model.getRep(tree.root)
148             X_predict.append(tree_rep)
149             Y_gold.append(tree.root.label)
150         acc = clf.score(np.array(X_predict),np.array(Y_gold))
151         return acc
```

REFERENCES

[1] Thomson Reuters. Accessed on Jan 2018. Open Calais. `http://www.opencalais.com/`.

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

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

[4] Supplementary material: Ideology detection in the indian mass media. `https://tinyurl.com/y4zc5k8d`.

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

[6] L. M. Bartels. Messages received: The political impact of media exposure. *American political science review*, pages 267–285, 1993.

[7] D. S. Benjamin and R. Bhuvaneswari. Bhoomi : ‘ e-governance ’ , or , an anti-politics machine necessary to globalize bangalore ? 2007.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

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

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

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

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

[13] S. Gerrish and D. Blei. Predicting legislative roll calls from text. pages 489–496, 2011.

[14] 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.

[15] J. Jagarlamudi, R. Udupa, and H. III. Incorporating lexical priors into topic models.

[16] A. K. Kanungo. Book review: Jean dreze, sense and solidarity: Jholawala economics for everyone. *International Studies*, 56(1):68–70, 2019.

[17] Y. Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882. `http://arxiv.org/abs/1408.5882`.

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

[19] S. Lai, L. Xu, K. Liu, and J. Zhao. Recurrent convolutional neural networks for text classification. In *AAAI*, 2015.

[20] S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[21] 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`.

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

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

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

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

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

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

[28] D. Niven. Objective evidence on media bias: Newspaper coverage of congressional party switchers. *Journalism & Mass Communication Quarterly - JOURNALISM MASS COMMUN*, 80:311–326, 2003.

[29] P. E. Oliver and D. J. Myers. How events enter the public sphere: Conflict, location, and sponsorship in local newspaper coverage of public events. *American journal of sociology*, 105(1):38–87, 1999.

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

[31] 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.

[32] D. Scheufele and D. Tewksbury. Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of Communication*, 57:9–20, 03 2007.

[33] J. C. Scott. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, 1998.

[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. pages 1–11, 2019.

[35] 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. pages 1–11, 01 2019.

[36] H. Shi, H. Zhou, J. Chen, and L. Li. On tree-based neural sentence modeling. *CoRR*, abs/1808.09644, 2018.

[37] P. Shoemaker, T. Vos, and D. Stephen. Journalists as gatekeepers. *The Handbook of Journalism Studies*, 2009.

[38] 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*, pages 91–101, 2013.

[39] 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.

[40] K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China, July 2015. Association for Computational Linguistics.

[41] R. Veeraraghavan. Dealing with the digital panopticon: The use and subversion of ict in an indian bureaucracy. volume 1, pages 248–255, 12 2013.

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

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

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

[45] R. Williams and D. Edge. The social shaping of technology. *Research Policy*, 25:865–899, 1996.

[46] H. C. Wu, R. Luk, K.-F. Wong, and K.-L. Kwok. Interpreting tf-idf term weights as making relevance decisions. 26, 2008.

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