Partisan Bias of Indian Media

Course: Data mining (CS-685A)

Instructor: Dr. Arnab Bhattacharya

Submitted by:

Appu B(18111008)

Pranjal Jain(18111050)

Lt Cdr Rahul Raj(18111053)

Lt Cdr Karan Basson(18111030)

Group No: 12

November 15,2018

Problem Statement

Journalistic objectivity may refer to fairness, disinterestedness, factuality, and nonpartisanship, but most often encompasses all of these qualities. Objectivity in journalism aims to help the audience make up their own mind about a story, providing the facts alone and then letting audiences interpret those on their own. To maintain objectivity in journalism, journalists should present the facts whether or not they like or agree with those facts. Objective reporting is meant to portray issues and events in a neutral and unbiased manner, regardless of the writer's opinion or personal beliefs

However, the reality is far from the above stated. The Media houses display a particular kind of bias called as political bias. Due to this political bias the news articles in any newspaper tend to favor the propaganda/ ideology of some political party. The schemes run by a political party are highlighted and praised and shortcomings go uncovered while only the failures of the other political parties are covered. This does not give the reader a fair opportunity to form their own opinion and only give them one side of the story.

In response to this increasing political bias in Indian Media, we have analyzed the news articles from different news agencies and conclusively proved that the news is not neutral and tried to infer the bias of four different media houses viz-a-viz two major political parties of India, Congress and BJP.

INDEX

1 Introduction

In order to solve such a problem, this project aims to use data mining techniques to create a system able to detect political bias in news articles by implementing coverage bias, sentiment analysis, gatekeeping bias. Initially, the project focuses on well known controversial topics for political parties. The political parties under consideration here are Congress and BJP. This work is divided in three phases. First, we create a dataset, by scrapping data from four different news sources. Data is then cleaned and organised. Second, we segregate the data among fifteen topics of interest using domain knowledge. For the third phase, we calculate coverage, sentiment score and similarity between articles of news agency to infer the bias. Finally, we describe our results by various plots and conclusions.

2 Data Collection and Preprocessing

2.1 Collection of data

Data is scraped from news websites namely, The Hindu, India Today, NDTV and Tribune from year 2004 to 2018. Scrapping is done using the python library “BS4- BeautifulSoup". The data source URLs are:

I. http://archives.ndtv.com/
II. https://www.indiatoday.in/archives/story/
III. https://www.thehindu.com/archive/
IV. https://www.tribuneindia.com/archive.aspx

Approximately 3.5 lakh articles were scraped from above sources and formulated a comprehensive dataset to query upon. An approximate duration of 120 hours per news website was taken for scrapping the articles from the webpage.

2.2 Data cleaning

The scraped raw data is cleaned by removing stop words, noise, irrelevant html tags and the date format is made uniform across the scraped data using the python library ‘DateTime’. The entire data is lower cased for capturing all relevant keyword matches. The final CSVs of news agencies are organised as {Date, Topic, Article}.

2.3 Segregation of topics

As per the domain knowledge, fifteen topics are queried from the dataset and separate CSV files are created. The queried topics are *’demonetisation*', '*beef ban*', '*Rafale jets*', '*swachh bharat*', '*GST*', '*FDI*', '*aadhar*', '*adarsh scam*', '*digital india*', '*coal scam*', '*chopper scam*', '*Karnataka election*', '*UP elections*', '*Sabarimala*' and '*2g scam*'.

3 Computing Bias

Three approaches are used to calculate the bias of the media which are explained as below: -

3.1 Coverage Bias

Coverage bias (also known as visibility bias) is when actors or issues are more or less visible in the news. The segregated data topic wise are queried for the frequency of coverage of keywords ‘congress’,’cong’ and ‘bjp’. Furthermore, the coverage is computed over all the topics aiming to find the average bias. The result is plotted on pie chart topic wise for each news agency.

3.2 Sentiment Bias

Statement bias (also known as tonality bias or presentation bias) is when media coverage is slanted towards or against particular actors or issues. Sentiment score of each article is computed using python library ‘TextBlob’. The range of score is between -1 to +1, depicting negative and positive sentiment of news article. The resulting score is appended to the topics csv for comparison. Average scores are computed topic and news agency wise. Topics are then segregated into ‘PRO BJP’ (topics which are in favour of ‘BJP’), ‘ANTI CONGRESS’ and ‘NEUTRAL’ using domain knowledge. The classified topics are then plotted and analysed whereby computing the sentiment bias.

3.3 Gatekeeping Bias

Gatekeeping bias (also known as selectivity or selection bias) is when stories are selected or deselected, sometimes on ideological grounds. It is sometimes also referred to as agenda bias, when the focus is on political actors and whether they are covered based on their preferred policy issues.

Articles of a topic of one of the news agencies are selected and similarity is computed with the same topic covered by other news agencies in the same time period. Cosine similarity is computed using library “sklearn – TFIDFVectorizer”. Below equation finds the similarity between two articles.

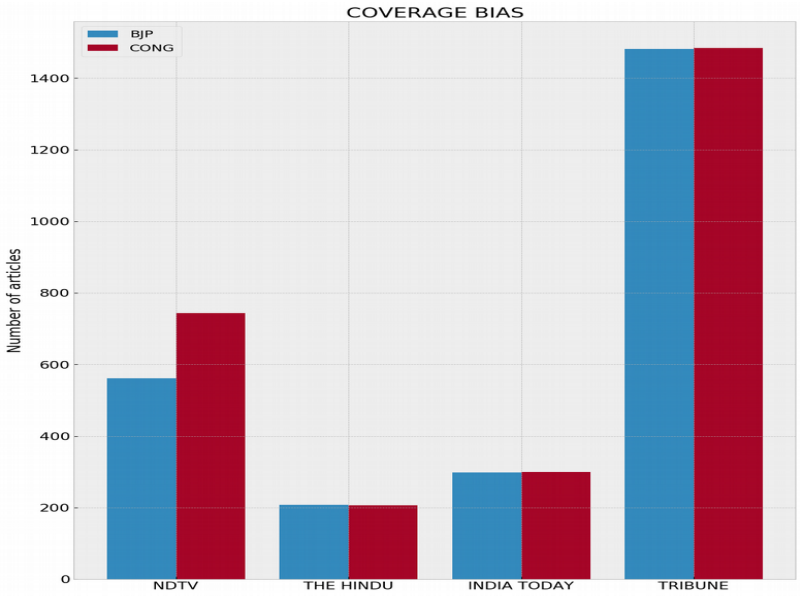
$$cosine(d_1, d_2) = \cos\theta(d_1, d_2) = \frac{d_1 . d_2}{|d_1||d_2|}$$

The resulting similarity matrix asserts the fact that the articles of a topic are presented or not by a news agency. Values in matrix are between 0 (completely dissimilar or independent) and 1 (identical).

4 Inference

4.1 Coverage bias:

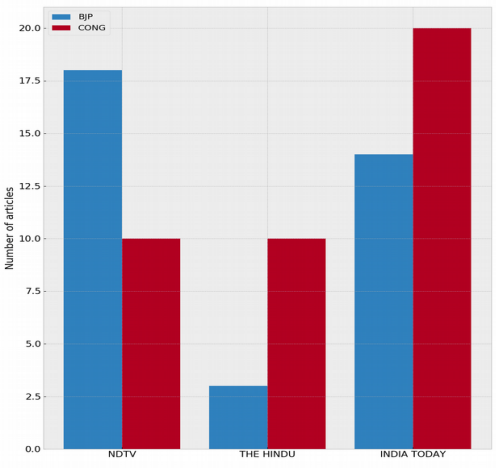
4.1.1 The overall coverage by all four sources across all topics is observed to be approximately equal. Plot of the same is as shown below: -



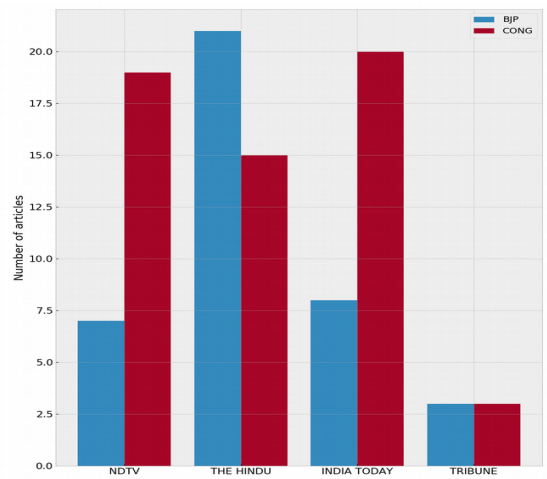
4.1.2 Topic wise news coverage

Below presented bar plots indicate the topic wise bias of news agencies. With analysis of topic wise plots, it is observed that NDTV has a coverage bias towards BJP while Tribune towards Congress. The bias of other news agencies could not be ascertained with coverage bias. It may be noted that all the plots are considered for articles published within common date range.

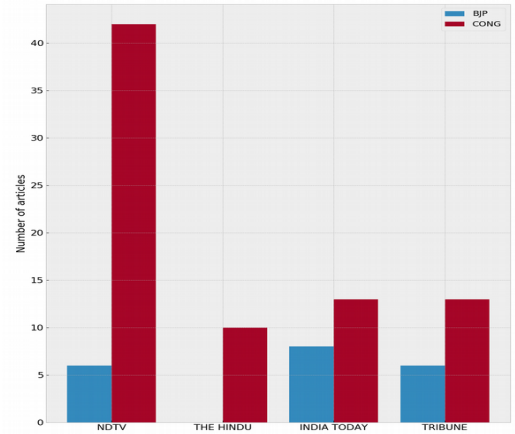
i) 2G Scam



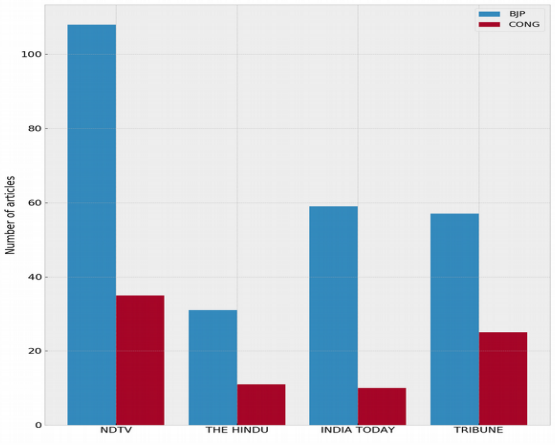
ii) Aadhar



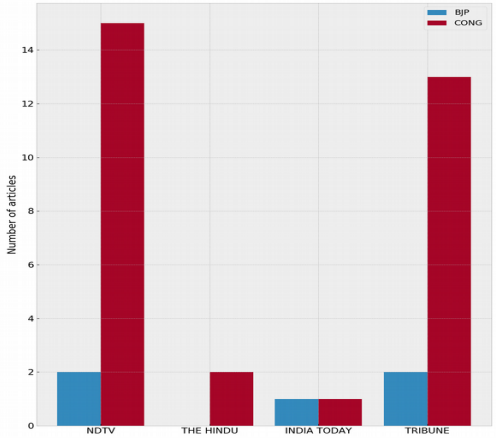
iii) Adarsh Scam



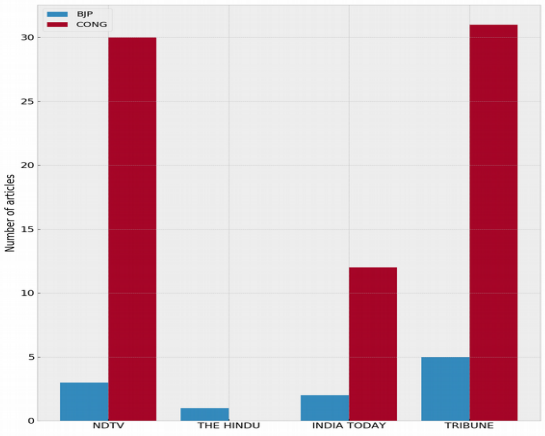
iv)Beef Ban



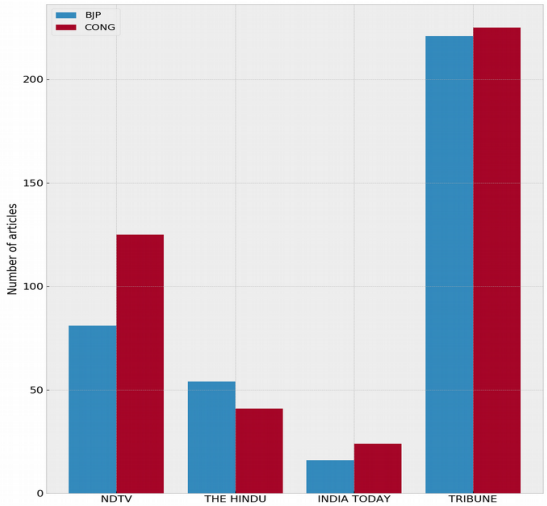
v) Chopper Scam



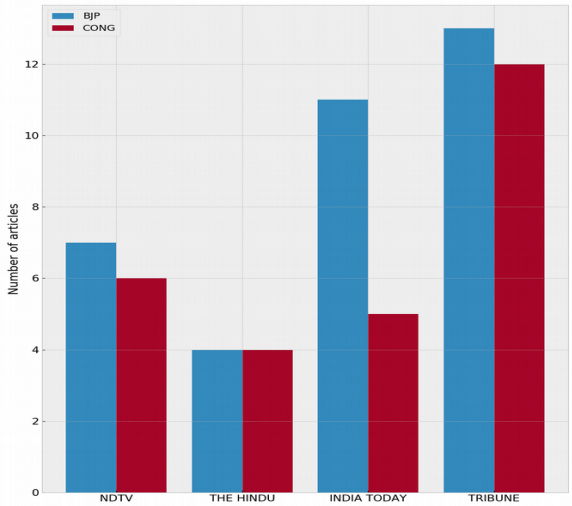
vi) Coal Scam



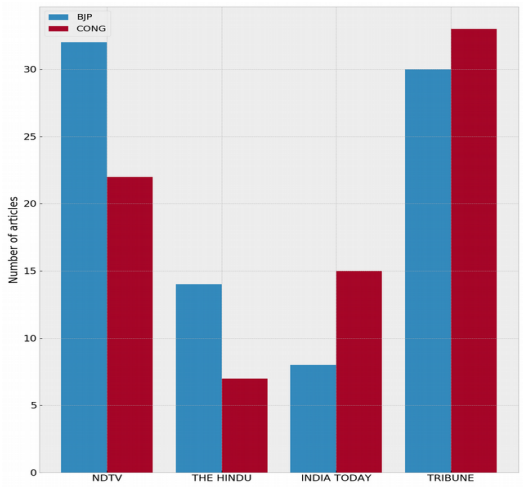
vii) Demonetisation



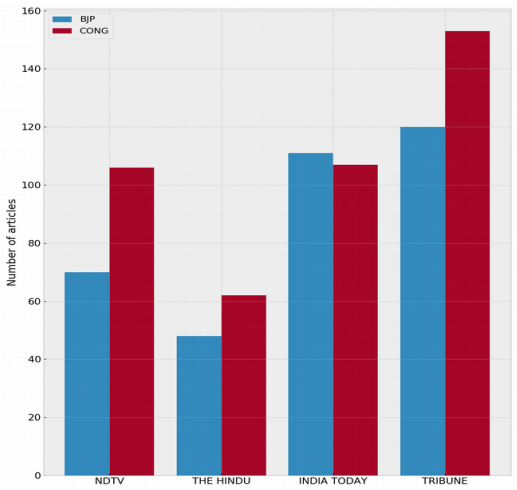
viii) Digital India



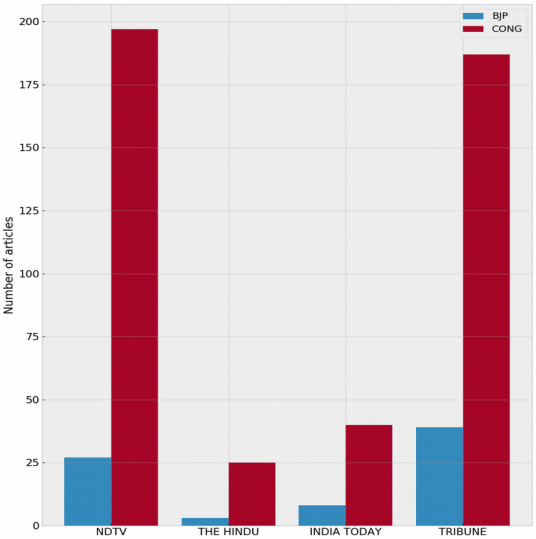
ix) Foreign Direct Investment



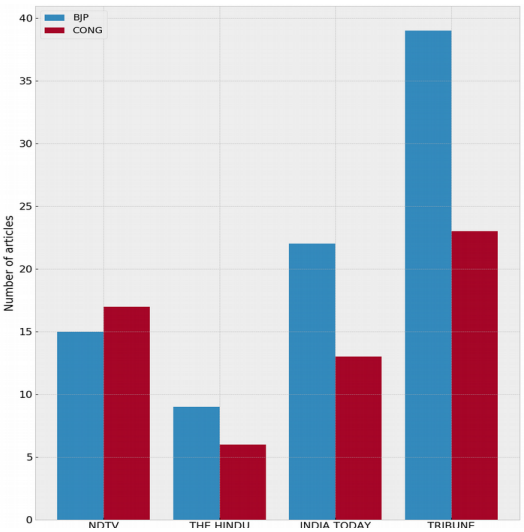
x) Goods and Services Tax



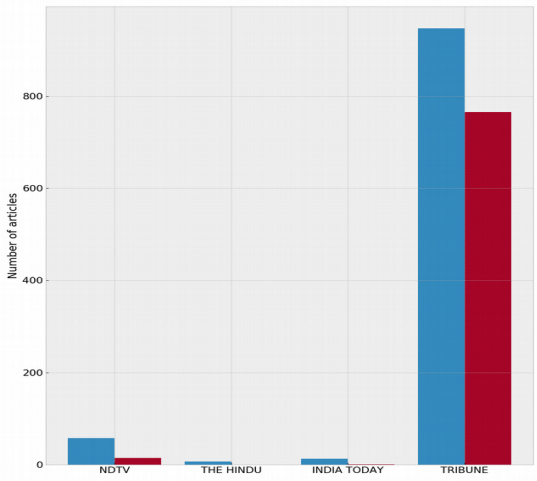
xi) Rafale



xii) Swachh Bharat



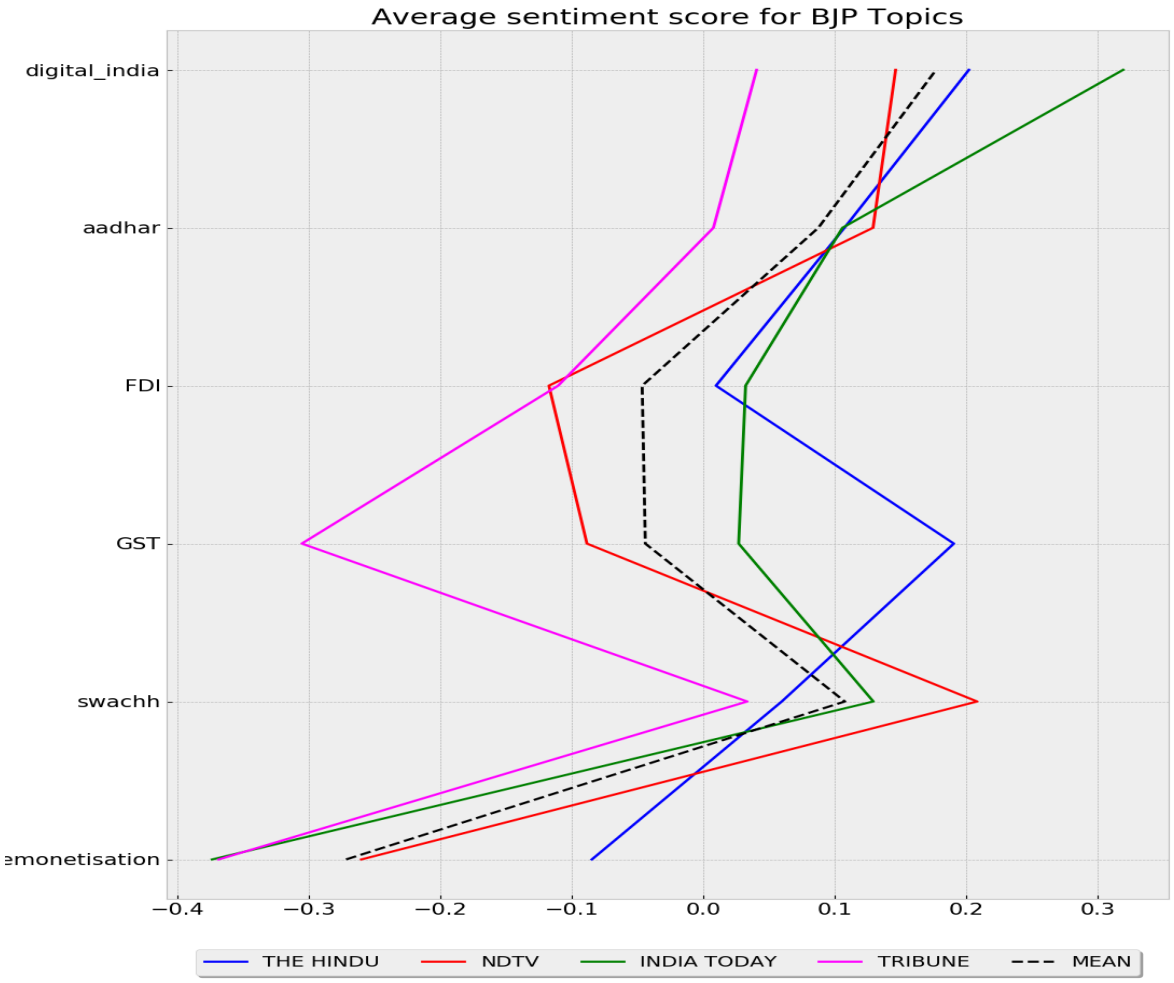
xiii) Uttar Pradesh Elections



4.2 Sentiment Bias

For sentiment analysis, we have grouped the topics under consideration into three categories, viz. BJP centric, Congress centric and Neutral Topics. The dotted line in the plot is the mean of the sentiment scores of each category. The goal is to compute how many standard deviation away the scores of news are from mean value. The inference from plots under each category are:

(i) BJP Centric Topics

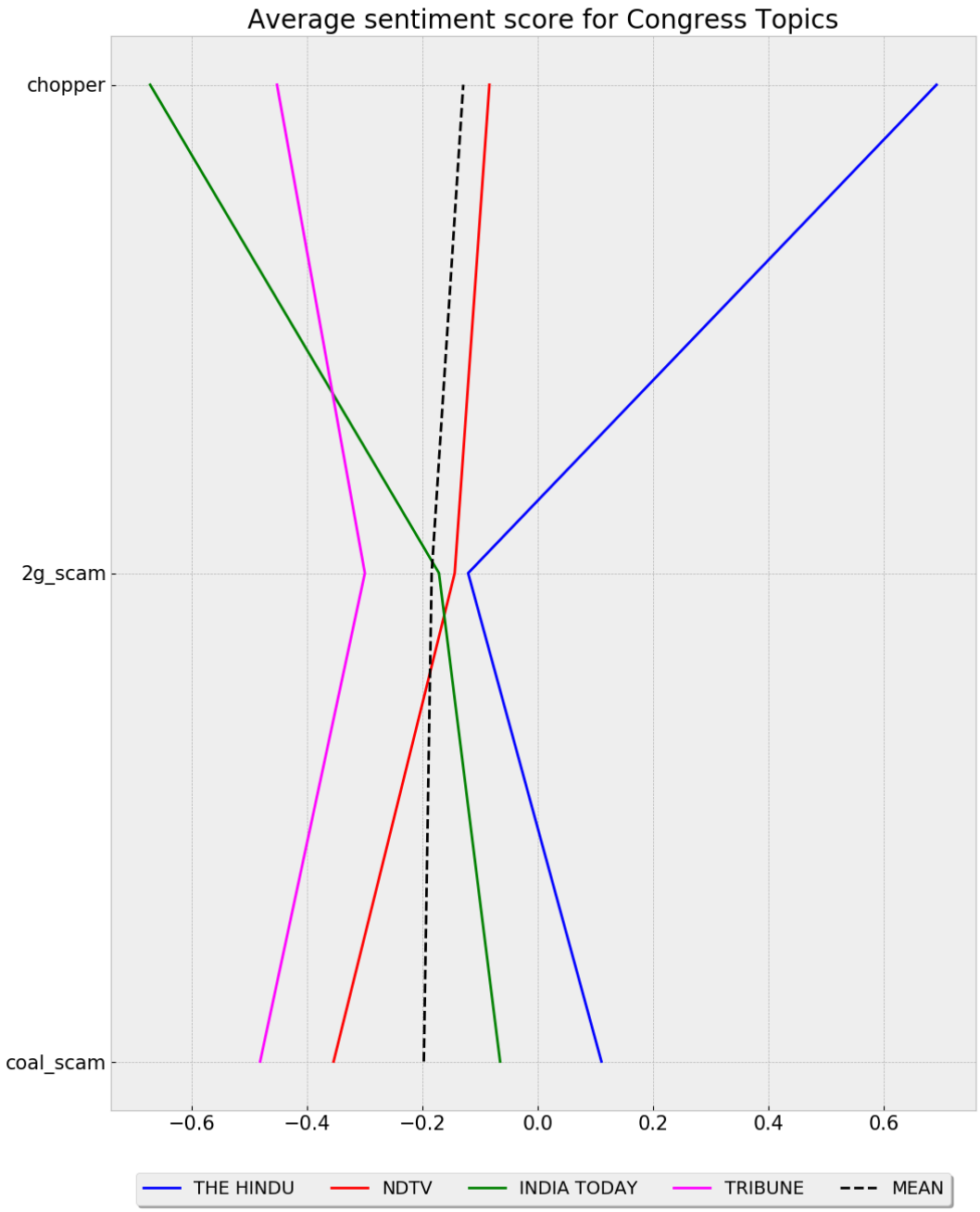


Inference from the above plot:-

- Hindu, India Today are observed as mostly above mean, NDTV is neutral and Tribune is below mean

- Clearly, Tribune is biased towards Congress as it has projected the pro BJP topics in negative sentiment consistnetly. Same inference was observed from the coverage bias results and thus ascetaining the bias.

(ii) Congress centric Topics

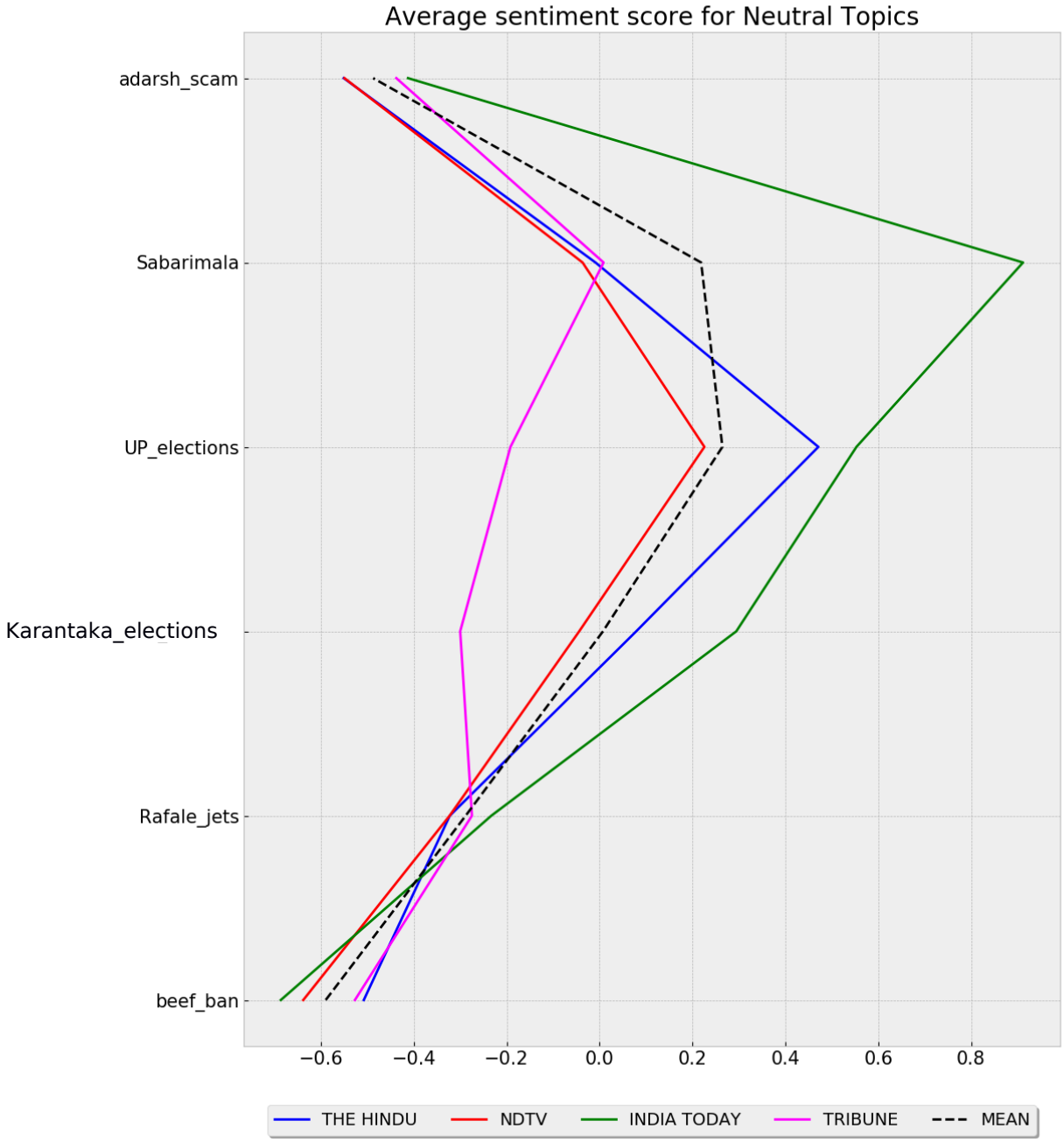


Inference from the above plot:-

◆ Tribune is observed to be below the mean score, NDTV neutral, Hindu above mean and, India Today could not be judged from data available.

◆ It is infered that Tribune is critic towards Government in Power. NDTV is observed to be in neutral reporting thus ascertaining the observation from previous plot.

(iii) Neutral Topics



Inference from the above plot:-

◆ Overall, India Today is showing positive bias and Tribune negetive bias towards neutral topic. NDTV and Hindu are on the neutral side.

◆ For topic ‘Sabarimala’, India Today has a high score above mean while other three media have reported with same sentiment. For topic ‘UP election’ and ‘Karnataka election’, all four media has their own sentiment. And thus a clear political bias, even though bias towards BJP or Congress is not ascertained.

◆ Table depicting mean, standard deviation and score of news agencies versus topics:

	Topics	Mean	Standard Deviation	The Hindu	Tribune	NDTV	India Today
0	demonetisation	-0.271493	0.117059	-0.084572	-0.368307	-0.259807	-0.373287
1	beef_ban	-0.589661	0.074760	-0.507646	-0.526739	-0.637969	-0.686289
2	Rafale_jets	-0.288147	0.037401	-0.322341	-0.274690	-0.322669	-0.232890
3	swachh	0.107780	0.067829	0.059785	0.033539	0.208376	0.129420
4	GST	-0.043940	0.180451	0.190632	-0.305125	-0.088320	0.027052
5	FDI	-0.046348	0.067898	0.009768	-0.110220	-0.117243	0.032303
6	aadhar	0.087626	0.046976	0.107485	0.007833	0.129201	0.105987
7	adarsh_scam	-0.487512	0.063186	-0.550839	-0.437548	-0.549292	-0.412368
8	digital_india	0.177127	0.100520	0.202092	0.040679	0.146375	0.319360
9	coal_scam	-0.197708	0.233196	0.110381	-0.481604	-0.354222	-0.065387
10	chopper	-0.129258	0.518511	0.691667	-0.452332	-0.083911	-0.672455
11	Karnataka_election	0.007425	0.214869	0.078998	-0.300000	-0.043807	0.294510
12	UP_elections	0.264510	0.289678	0.471043	-0.191827	0.225778	0.553046
13	Sabarimala	0.218754	0.400431	-0.008776	0.008641	-0.036605	0.911758
14	2g_scam	-0.184044	0.069302	-0.120760	-0.300000	-0.144045	-0.171373

4.3 Gatekeeping Bias

Similarity of topics wise articles of all news agnecies is computed using tf-idf and threshold assumed for similarity is 50%. Table below depicts the gatekeeping bias of news agencies:

Demonetisation

Total Articles (H - 357, IT - 74, NDTV - 876 , TRIB - 809)

	THE HINDU	INDIA TODAY	NDTV	TRIBUNE
THE HINDU	NaN	32	67	38
INDIA TODAY	NaN	NaN	19	23
NDTV	NaN	NaN	NaN	248
TRIBUNE	NaN	NaN	NaN	NaN

Beef Ban

Total Articles (H - 97, IT - 115, NDTV - 291 , TRIB - 112)

	THE HINDU	INDIA TODAY	NDTV	TRIBUNE
THE HINDU	NaN	10	28	7
INDIA TODAY	NaN	NaN	21	2
NDTV	NaN	NaN	NaN	6
TRIBUNE	NaN	NaN	NaN	NaN

Rafale Jets

Total Articles (H - 42, IT - 64, NDTV - 298 , TRIB - 388)

	THE HINDU	INDIA TODAY	NDTV	TRIBUNE
THE HINDU	NaN	10	14	11
INDIA TODAY	NaN	NaN	31	27
NDTV	NaN	NaN	NaN	163
TRIBUNE	NaN	NaN	NaN	NaN

Swachh Bharat

Total Articles (H - 208, IT - 135, NDTV - 334 , TRIB - 245)

	THE HINDU	INDIA TODAY	NDTV	TRIBUNE
THE HINDU	NaN	8	10	20
INDIA TODAY	NaN	NaN	14	22
NDTV	NaN	NaN	NaN	61
TRIBUNE	NaN	NaN	NaN	NaN

GST

Total Articles (H - 1226, IT - 732, NDTV - 978 , TRIB - 574)

	THE HINDU	INDIA TODAY	NDTV	TRIBUNE
THE HINDU	NaN	74	96	53
INDIA TODAY	NaN	NaN	127	125
NDTV	NaN	NaN	NaN	252
TRIBUNE	NaN	NaN	NaN	NaN

FDI

Total Articles (H - 157, IT - 123, NDTV - 607 , TRIB - 231)

	THE HINDU	INDIA TODAY	NDTV	TRIBUNE
THE HINDU	NaN	30	33	47
INDIA TODAY	NaN	NaN	17	30
NDTV	NaN	NaN	NaN	71
TRIBUNE	NaN	NaN	NaN	NaN

Aadhar

Total Articles (H - 346, IT - 132, NDTV - 224 , TRIB - 30)

	THE HINDU	INDIA TODAY	NDTV	TRIBUNE
THE HINDU	NaN	24	13	5
INDIA TODAY	NaN	NaN	13	5
NDTV	NaN	NaN	NaN	5
TRIBUNE	NaN	NaN	NaN	NaN

Adarsh Scam

Total Articles (H - 27, IT - 37, NDTV - 120 , TRIB - 47)

	THE HINDU	INDIA TODAY	NDTV	TRIBUNE
THE HINDU	NaN	14	10	8
INDIA TODAY	NaN	NaN	24	22
NDTV	NaN	NaN	NaN	57
TRIBUNE	NaN	NaN	NaN	NaN

Inference:

- The above table represent similar articles identified across various media. The Gatekeeping bias is evident from the above data and confirms the selective projection/ reporting of news.

5 Conclusion

It is evident from the above study that Indian Media is politically biased. However, a single technique cannot ascertain the political inclination. Thus calculating the bias from 'coverage', 'sentiment', and 'gatekeeping', we have cross verified the results to establish following facts: -

(a) NDTV has a neutral political view in objective presentation.

(b) Tribune is a political critic of Government in Power.

(c) All four media has got individual political stand during election period. (Evident from analysis of topics '*UP Election*', '*Karnataka Elections*')

(d) Hindu and India Today is having inclination towards Government in Power.