Analysing the political content uploaded on YouTube near Indian Elections

This report is submitted as the fulfilment of VII.3.2 Making Society Smart through Computational Social Systems requirements of B. Tech (IT & Mathematical Innovations)
At Cluster Innovation Center, the University of Delhi

Submitted By

Nikunj Saini (20312915028)

Saransh Chopra (20312915047)

Under the Supervision of

Dr. Sachin Kumar

Cluster Innovation Center
University of Delhi, Delhi
India-110007

Acknowledgment

We would like to express our special thanks to our supervisor Dr. Sachin Kumar who allowed us to work on this research project titled “Analysing the political content uploaded on YouTube near Indian Elections”. Without his help, it would have been very difficult to complete this project. We came to know about so many new things and had a good experience. We would also like to show gratitude to our parents for their kind support at home.

Nikunj Saini (20312915028)

Saransh Chopra (20312915047)

Table of Contents

Abstract

Most people in India rely on the media for current events and news. Media holds a key role in a democracy and is often termed the 'Fourth Pillar of Democracy'. Public opinion can be shaped by the media, and social change can be sparked by it. It can advocate for human rights, draw attention to pressing concerns, and inspire people and communities to tackle societal difficulties. Because the media holds such an important role in today's society, biased media can be harmful to society. Along with television, media has a strong presence on YouTube, a social media application that is comparatively understudied in India's context. YouTube videos have in the past caused extreme outrage by instigating a group of citizens, especially during elections. This work aims to analyze the trends of the content posted on YouTube at the time of the General Indian Elections of 2019 and Karnataka State Elections of 2023. This work collects an original continuous dataset of videos posted by the top 10 (TRP) English channels in the period of 3 months before elections and label the data as pre-government, anti-government, or neutral using a novel mathematical model. Furthermore, we analyze monthly trends in the data and see how they compare to each other, and to different elections. Finally, we fine-tune textual models on this dataset to automatically classify any YouTube video as pro-government, anti-government, or neutral. As an extension, we train a Large Language Model on a subset of the dataset to assess if the model generates text inclined to a specific political party or stays neutral.

Objectives

1. An original dataset of YouTube videos posted 3 months before the Indian General Elections of 2019 and Karnataka State Elections of 2023.
2. A word-frequency-based mathematical model for manually labeling a YouTube video as pro-government, anti-government, or neutral based on a score.
3. Analyzing the trends in the dataset and how these trends change as the election date approaches.
4. Deep Learning textual models trained on the dataset to classify any video as pro-government, anti-government, or neutral.
5. A Large Language Model trained on a subset of data to assess whether it exhibits a bias toward a specific political party or remains neutral.

Introduction

With an active user count of more than 574 million, India ranks as the number 1 country in terms of number of active users on YouTube. YouTube serves as a medium for both entertaining and official content, through influential and news channels. Combining the ability to rewatch and share a news section with the ease of accessing content, YouTube videos are destined to spread more rapidly than television content, making them the perfect medium for news channels to spread polarising content. These news channels can be biased in many ways, including, but not limited to politically biased, economically biased, culturally biased, or conformationally biased. The public's sense of reality can be influenced by media bias, which can result in misinterpretations of significant topics and events. The public may be less informed and more divided when media outlets persistently present one side of an issue. Trust in media organizations is damaged by bias. It becomes difficult to separate fact from fiction when consumers believe that media sources are politicized or unreliable. YouTube, being more accessible than traditional television, spreads this polarising content at an even higher speed.

The public's sense of reality can be influenced by media bias, which can result in misinterpretations of significant topics and events. The public may be less informed and more divided when media outlets persistently present one side of an issue. Trust in media organizations is damaged by bias. It becomes difficult to separate fact from fiction when consumers believe that media sources are politicized or unreliable. YouTube, being more accessible than traditional television, spreads this polarizing content at an even higher speed.

The amount of polarizing content on every social media rapidly rises near the election dates, and this rise manipulates voters and their votes. There have been little to no studies on this polarizing content in the context of the Indian population and politics. To shed light on the role of YouTube videos in shaping India's vote bank, this study aims to collect data from YouTube spread evenly around the central election of 2019 and the Karnataka state election of 2023. The data will be collected from 3 months before the election date, in this case, from January to April.

In this research, we seek to address critical gaps in our understanding of the impact of YouTube and social media on the Indian electoral process. By analyzing content trends and biases in the lead-up to elections over multiple years, the study aims to provide valuable insights into the potential influence of digital media on Indian voters and the democratic process as a whole. The findings of this research may have significant implications for media regulation, voter education, and the overall health of Indian democracy.

Literature Survey

The media's ability to influence a democratic society has made worries about mainstream media bias a contentious and important topic in journalism. Strong evidence of the impact of media bias on voters has been established by the study. Voters specifically adjust their party preferences according to the tone of the media that they are exposed to. Additionally, they would assess parties more favourably if they received greater media attention for addressing their own preferred issues [5]. Evidence of media bias effects on candidate trait perceptions, particularly on more politically significant qualities, has been discovered in the study [6]. Voters specifically adjust their evaluations of a candidate's assertiveness and ability in response to negative media representations of the candidate. If politicians can address their own preferred topics comparatively more prominently in media coverage, voters will also view those candidates' political qualities and general image more favourably, likely priming the precise problems on which they want their political acumen to be evaluated [6].
A similar approach was followed to create system-wide bias metrics based on the frequency with which media sources cite bias in mainstream and social media to measure prejudice in these contexts [7]. They had an observation that the popularity distribution on social media is more severely distorted and unequal than on other platforms because it is more social, i.e., more influenced by network and external influences. Though many of the voices seem to be echoes, there may be more than ever.

With a million internet users, a thousand TV channels, 100,000 registered publications, several hundred languages and dialects, and a growing population, India has become a challenging place for the government to "regulate" the media, despite their best efforts. But in a democracy, when the press is meant to be unrestricted, the governing party uses different strategies to control the media. Numerous media outlets are taken over by business magnates who have ties to the ruling party; political parties hire digital teams to produce trolls and fake news. The nation's politics have fundamentally become divisive. Digital new media technologies aggravate this division [8].

Research Questions

- How much pro-government, anti-government, or neutral content is uploaded on YouTube during the election year, and how does its quantity vary near the election period?
- Are citizens interacting more with a particular type of content uploaded on YouTube near elections or are citizens consuming content wisely?
- Is there a particular news channel which is constantly favouring or opposing the current government, or are they all uploading neutral content?
- Can we label any Indian news YouTube video as pro-government, anti-government, or neutral?
- Will a Large Language Model trained on a subset of these transcripts be biased towards a party or will it stay neutral?

Dataset and Description

This study collected a novel dataset encompassing videos uploaded during the three months leading up to each of the selected election years (i.e., 2019 and 2023) from eight English YouTube channels of registered Indian media houses. Table 1 lists these 8 YouTube channels along with their lifetime YouTube statistics.

YouTube Channel	Subscribers	Views (lifetime)	Total videos (lifetime)
Republic World	5,820,000	2,224,770,476	103,881
Times Now	3,730,000	1,735,635,668	92,628
CNN News 18	5,580,000	2,358,876,403	145,879
DD India	386,000	7,387,957	2,841
India Today	7,900,000	2,891,500,097	160,357
WION	8,140,000	4,412,210,258	121,508
Mirror News	656,000	281,272,299	69,478
NewsX	911.000	259,545,072	95,916

Table 1: Statistics of YouTube media channels used in this study.

These YouTube channels were selected from the internationally recognized TRP rankings published by BARC India. Table 2 lists down the data and metadata attributes collected for an individual video.

Data attribute	**Description**
Unique ID	Unique YouTube ID of the video.
Title	Title of the video.
Description	Description of the video.
Publish date	Date on which the video was uploaded.
Publishing channel	Channel on which the video was uploaded.
Thumbnail URL	YouTube URL of video's thumbnail.
Transcripts	Entire transcript (captions) of the video in JSON and textual format.
Source of generation	Source which generated these transcripts (can be manually uploaded or automatically generated by YouTube).
Original language	Original language of the transcripts of the video.
Views	Numer of views on the video.
Comments	Root-level comments and replies to those comments on a video.
Likes	Number of likes on a video.

Table 2: Attributes of a video used in this study.

The entirety of the videos uploaded by media on their YouTube channels did not include a manually uploaded transcript; therefore, all the collected videos' transcripts are YouTube generated. A few videos had a non-English language ('de', 'fr', 'es', 'pt', 'ru', 'nl', 'it', 'hi', 'id', 'ko', 'tr',') set as their transcript language; hence we translated these transcripts to English ('en') using Google Translate. Finally, most uploaded videos had their transcripts disabled; thus, this study consists of a subset of videos posted by these media channels, a subset that has transcripts enabled on YouTube.

Most data points were collected using YouTube's official API (V3), but the transcripts were specifically extracted using a free-to-use and open-source Python library called youtube_transcript_api. Lastly, the extracted non-political videos were manually removed from the dataset. Tables 3 and 4 show the distribution and the total number of videos collected for studying Indian General Elections and Karnataka State Elections respectively.

Channel name	**Videos collected**
Republic World	124
Times Now	107
CNN News 18	105
DD India	127
India Today	168
WION	310
Mirror News	204
NewsX	119
Total	**1264**

Table 3: Distribution of videos collected - General Elections.

Channel name	**Videos collected**
Republic World	105
Times Now	103
CNN News 18	87
DD India	126
India Today	122
WION	330
Mirror News	199

NewsX	100
Total	**1172**

Table 4: Distribution of videos collected - Karnataka State Elections.

Used Methods and Approaches

Data labelling using statistical model

The labelling of video transcripts was carried out with the help of a statistical model. This model is based on the frequency of keywords that are used in a given transcript. We manually selected keywords relevant to Indian politics in the given period for the model. There are broadly 5 categories of words that are used in a news video, as listed in table 5.

Keyword Type	**Description**
Government references	Words that refer to the people related to the government and schemes launched by the government. They may also include references to governmental organizations as well.
Opposition references	Words that refer to the people related to opposition parties. These may include alleged scams that the government might be involved in and the government schemes that they oppose.
Appreciating words	Words that appreciate a person, organization or scheme are included in this category. They may be used about the government or the opposition.
Criticizing words	Words that are either questioning or criticizing a person, organization or scheme will be counted in this category.

Table 5: Types of keyword in a political youTube vide;.

The novel score is calculated as the ratio of the product of government references and appreciating words to the product of criticizing words and opposition words. To account for the fact that this score does not count that the news is also appreciating the opposition fails if the news channel is criticizing the government, and does not consider the fact that the news video is both criticizing the opposition and appreciating the government with almost the same frequency, we use a log-biased score. The bias score θ, for a given video, can be calculated as

$$\theta = log\left[\left(\frac{f_{gr} \cdot f_{aw} + f_{nt}}{f_{gr} \cdot f_{cw} + f_{nt}}\right)\left(\frac{f_{or} \cdot f_{cw} + f_{nt}}{f_{or} \cdot f_{aw} + f_{nt}}\right)\right]$$

$f_{gr} -$ $frequency\ of\ words\ belonging\ to\ government\ references$

$f_{or} -$ $frequency\ of\ words\ belonging\ to\ opposition\ references$

$f_{aw} -$ $frequency\ of\ words\ belonging\ to\ appreciating\ words$

$f_{cw} -$ $frequency\ of\ words\ belonging\ to\ criticizing\ words$

$f_{nt} -$ $frequency\ of\ words\ that\ do\ not\ belong\ to\ any\ catagory$

The above formula solves the issues enlisted above. Given that headlines contain news from every domain, classifying them as pro-government or anti-government does not make any sense; hence, the margin of the bias label can be adjusted as and when required.

Automatic classification using textual Deep Learning models

This study used the BERT (bert-base-uncased) [9] and the RoBERTa [10] model to create an automatic framweorkd for labelling any political YouTube video as pro-government, anti-government, or neutral. The dataset was divided in a ratio of 80:20 and models were prepared with the hyperparameters and configurations listed in Table 6 and Table 7 respectively.

Hyperparameter or Configuration	Value
Batch size	16
Optimiser	AdamW
Weight decay	4e-3
Learning rate	1e-5
Epochs	10
Learning rate scheduler	Linear

Table 6: Hyperparameter and configuration values for the trained BERT model.

Hyperparameter or Configuration	Value
Batch size	16

Optimiser	AdamW
Weight decay	1e-2
Learning rate	2e-5
Epochs	20
Learning rate scheduler	None

Table 7: Hyperparameter and configuration values for the trained RoBERTa model.

Hyperparameter or Configuration	**Value**
Batch size	16
Optimiser	AdamW
Weight decay	1e-2
Learning rate	2e-5
Epochs	20
Learning rate scheduler	None

Table 8: Hyperparameter and configuration values for the trained BART model.

Hyperparameter or Configuration	**Value**
Batch size	16
Optimiser	AdamW
Weight decay	1e-2
Learning rate	2e-5
Epochs	20
Learning rate scheduler	None

Table 9: Hyperparameter and configuration values for the trained Distilled BERT model.

Large Language Model

Given that we did not have enough resources to train a Large Language Model, we decided to use the Large Language Model on the internet. A subset of the transcripts was converted to a PDF and was fed to ChatPDF [11]. This subset was small enough to fit in the free tier version of ChatPDF, focussing on the Government's Rafale deal. ChatPDF's configuration and hyperparameters are closed source; hence, we could only play with the Large Language Model by asking it to form opinions on the input data.

Research Methodology

This study collected a novel dataset of YouTube video data and metadata in a three-month time period from the Indian General Elections of 2019 and the Karnataka State Elections of 2023. Further work is carried out solely on the transcripts of the videos as the other data points are either too few in quantity or show no correlation with the type of uploaded content. The data was collected using the YouTube (V3) API.

Data preprocessing steps played a crucial role in ensuring the quality and relevance of the dataset. The transcripts were pre-processed to make them ready for Deep Learning algorithms and our statistical model. This will involve content categorization, where videos will be categorized into one of three groups: pro-government, anti-government, or neutral. Additionally, data cleaning procedures were applied to remove duplicate videos and irrelevant content that does not pertain to Indian elections, thus ensuring the dataset's integrity.

The mathematical model paved the way for the Deep Learning model and data analysis. The model. The sole idea behind the mathematical model was based on how frequently a particular political party, individual or organisation is referred to in a video and in what context it is being said. Some adaptations were made to overcome the drawbacks of the first model, and the adapted model worked seamlessly and gave us accurate results up to a great extent.

Data analysis in this research encompassed various techniques and approaches. Content analysis will involve a thorough examination of the textual and numerical content within the collected videos to assess the nature of the uploaded content and its interaction with the consumers. Temporal trends and word clouds helped identify patterns and shifts in the type of content and its dissemination leading up to elections.

The Deep Learning model created a unified framework for the future capable of labelling any political YouTube video as pro-government, anti-government, or neutral with decent accuracy. An optional component of this research involved utilizing a Large Language Model like to answer questions after being trained on the political dataset. This helped us determine if there is an abundance of a particular type of content on YouTube or if the media is covering everything neutrally.

The ultimate goal of this research was to provide meaningful insights into the role of YouTube content in shaping the Indian electoral landscape. Findings are summarized and presented in this comprehensive research report, including data visualizations and statistical analyses.

Data pre-processing

The data was manually combed to remove any non-political videos from all the individual channel datasets. Furthermore, the transcripts were cleaned by removing stop words, punctuation marks, and other special characters irrelevant to the videos. The transcripts were also lemmatized using along with extracting the Named Entities from them. The language preprocessing work was carried out through spaCy's en_core_web_trf model.

This process also included labelling the data as pro-government, anti-government and neutral. This was done by calculating the slant score of each entity through the statistical formula described in the section above. If the slant score of the entity is greater than 0, then it is most likely to favour the government, be it by either criticizing the opposition or admiring the works done by the government. If the score is less than 0, then it is more likely to be talking about the flaws of the government or appreciating the opposition. If the score comes out to be zero, then it is most likely to be a news headline or a politically neutral entity.

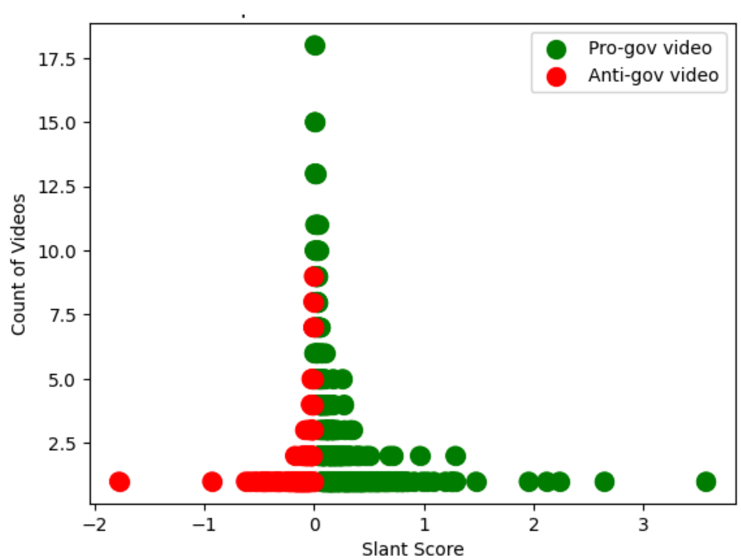

Figure 1: Slant score of videos (Karnataka State Elections) calculated using our mathematical model.

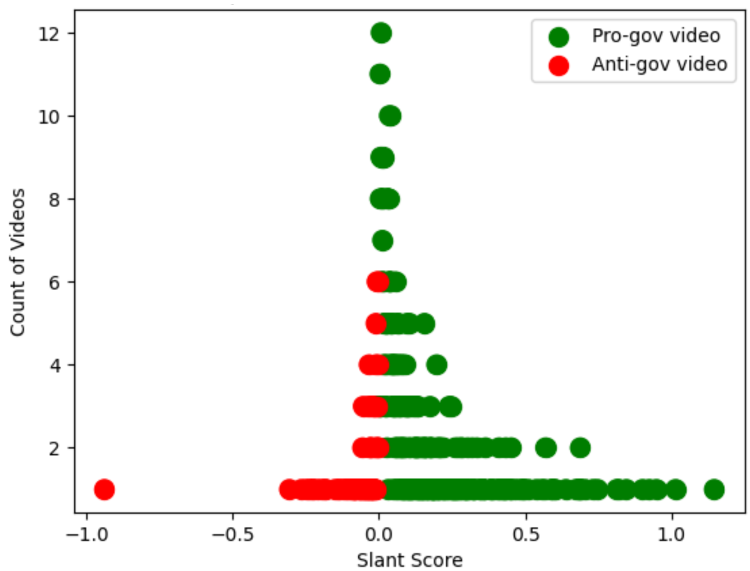

Figure 2: Slant score of videos (Indian General Elections) calculated using our mathematical model.

On further analysis, we found that the entities having a slant score within the range of -0.01 and 0.01 have headlines or general news clippings, which can neither be classified as pro-government nor anti-government. So, to label the entities, we considered all the entities that have a score of more than 0.01 to be pro-government and if the score is less than -0.01, then it is labelled as anti-government video. Rest all the videos whose absolute score is less than 0.01 are labelled as neutral videos.

Figure 3 and 4 show the distribution of the videos labelled with our statistical models. One can infer that the ProGov and Neutral videos account for more than 75% of the dataset whereas the AntiGov videos are very few.

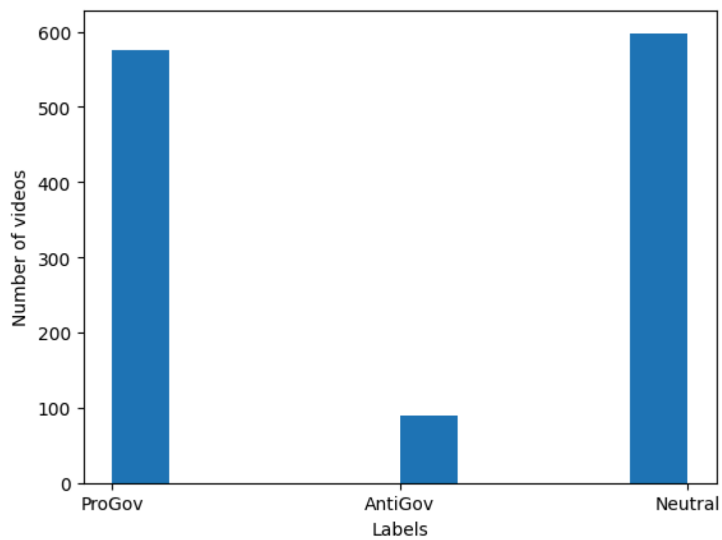

Figure 3: Distribution of videos (Indian General Elections) classified using our mathematical model.

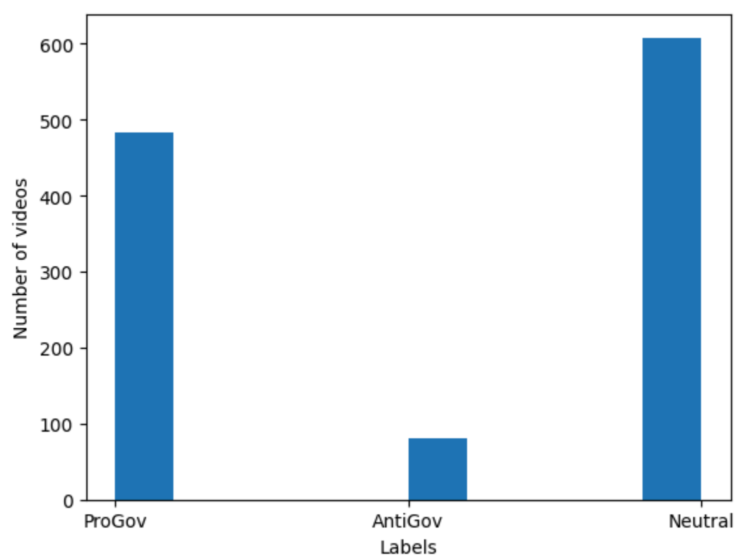

Figure 4: Distribution of videos (Karnataka State Elections) classified using our mathematical model.

Exploratory Data Analysis

We carried out extensive data analysis to figure out mathematical trends in both datasets. This analysis included plotting bar charts, custom frequency charts, and violin charts, as required by different data types. The plots were generated for both the elections independently and then compared to each other. The section below is divided into two halves, each of which describes the plots for a specific election period.

Indian General Elections 2019

Figure 5 qualitatively shows the type of content uploaded by specific YouTube channels before the Indian General Elections of 2019. This plot was generated by adding some noise to the data, given that the points would have overlapped each other perfectly if there was no noise. The figure shows that the amount of Neutral and ProGov content on YouTube was much more than the AntiGov content.

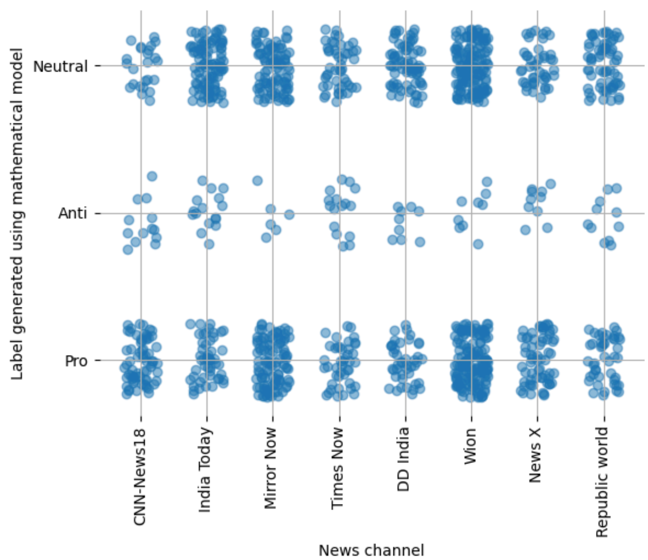

Figure 5: Qualitative distribution of videos (Indian General Elections) per YouTube channel.

Figure 6 gives a more quantitative version of figure 5. The figure shows that the number of AntiGov videos uploaded by each channel was less than 20, but at the same time, the number of ProGov and Neutral videos touched as high as 150. Moreover, we can see that the number of Neutral videos uploaded by India Today, DD India, Times Now, and Republic World was greater than the number of ProGov videos uploaded by the same channels. On the other hand, the number of ProGov videos uploaded by CNN News 18, Wion, Mirror Now, and NewsX were greater than the number of Neutral videos uploaded by the same channels.

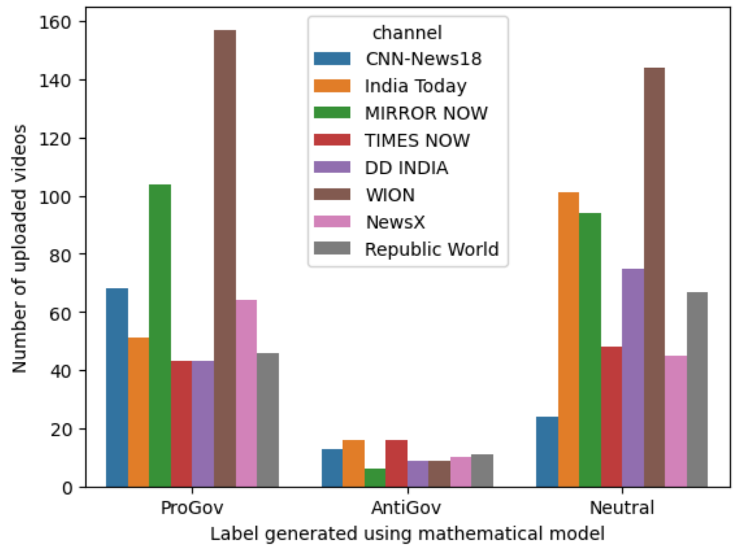

Figure 6: Quantitative distribution of videos (Indian General Elections) per YouTube channel.

Next, we analyze the number of likes, comments, and views on the videos. Given that the news channel use live streams to present current news and upload snippets of the live streams as YouTube video, citizens tend to comment less on these snippets. Similarly, the number of likes is low on these videos as compared to the livestreams. Figure 7 shows the number of likes plotted as a violin chart against the type of video. From the width of the violin plot, we can infer that the Neutral videos are gaining more likes than the ProGov and AntiGov videos. Analyzing further, the number of likes in ProGov and AntiGov videos looks alike but the ProGov videos have a slightly higher number of likes. We must also consider that these are an absolute number of likes and the number of ProGov, AntiGov, and Neural videos are not the same in our dataset.

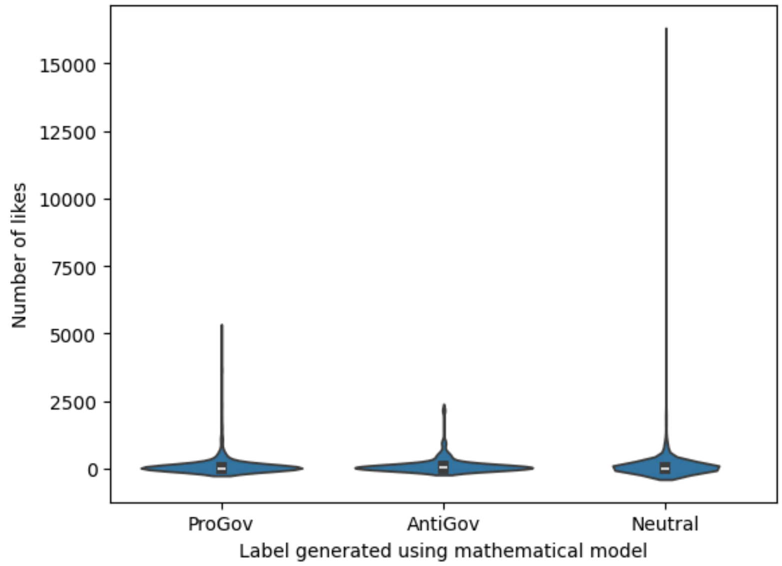

Figure 7: Numer of likes v/s type of uploaded videos (Indian General Elections).

Similarly, through figure 8, we can infer that the PrGov videos are gaining much more traction in terms of comments than the AntiGov and Neutral videos. The plot for the Neutral videos is extremely flat at 0, showing a scarcity of comments.

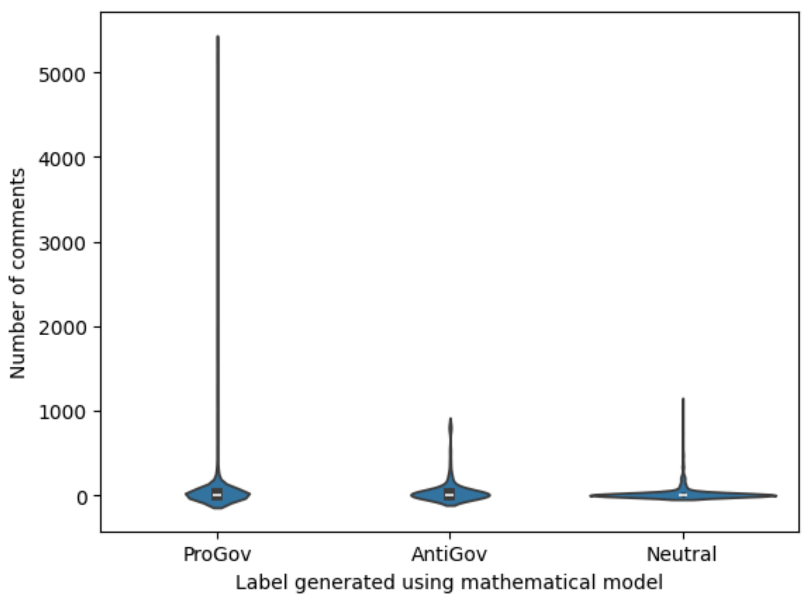

Figure 8: Numer of comments v/s type of videos (Indian General Elections).

In terms of views, in figure 9 every type of video seems to be gaining an equal distribution of views. The number of views is slightly higher for the ProGov videos and the lowest for AntiGov videos.

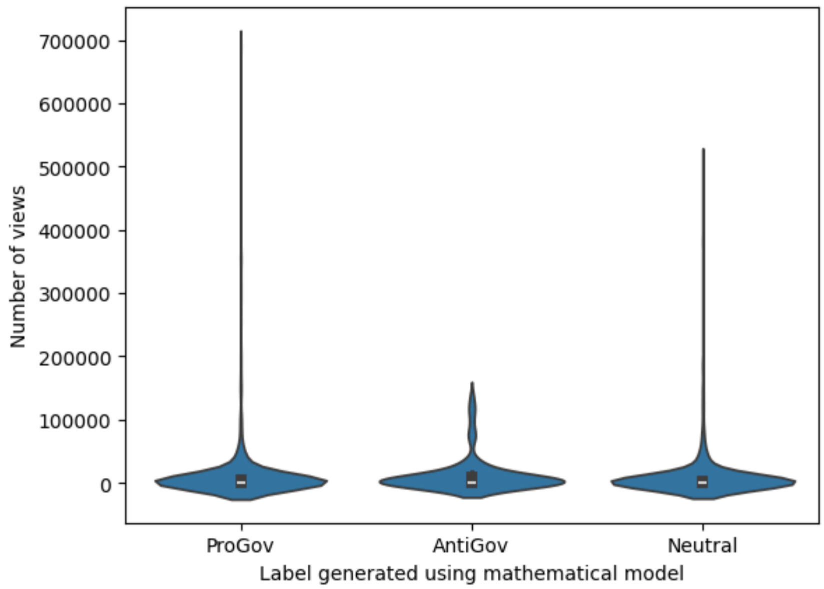

Figure 9: Numer of views v/s type of (Indian General Elections).

Lastly, we see the temporal trend in the number of different types of videos uploaded collectively by all the channels. Figure 10 shows a marginal trend of the number of uploads increasing before the elections and then decreasing back to normal as the election date approaches. More specifically, we see a peach around 1 month before the election date. There are also days when the frequency of uploads is comparatively low, even during the peak period.

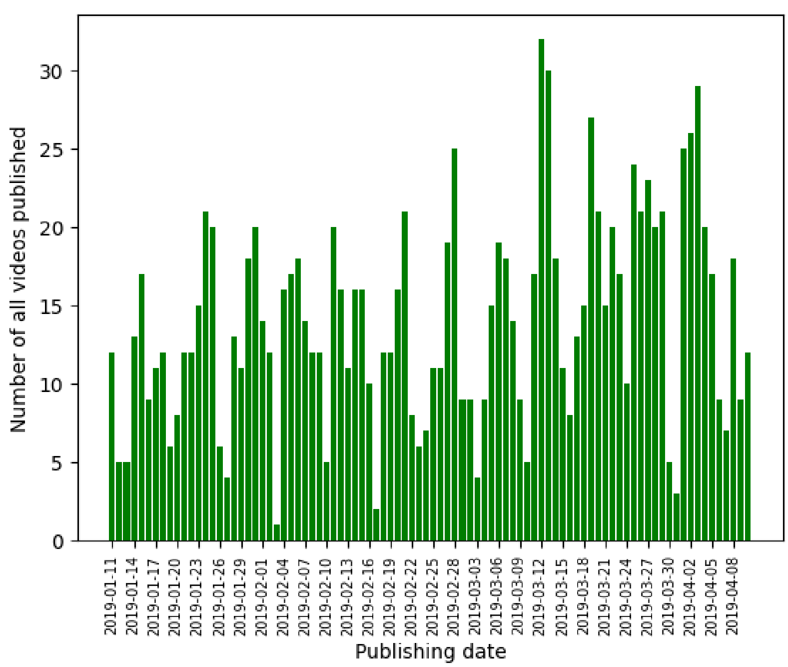

Figure 10: Temporal trend of all videos (Indian General Elections) by all YouTube channels.

Figure 11 shows a similar trend but only for the Neutral videos. We can see that this plot is more evenly spread out throughout the three months. Additionally, we see multiple peaks in this graph, one of which is one month before the election date. The Neutral uploads stay at an all-time high even after the total number of videos starts declining after the one-month peak.

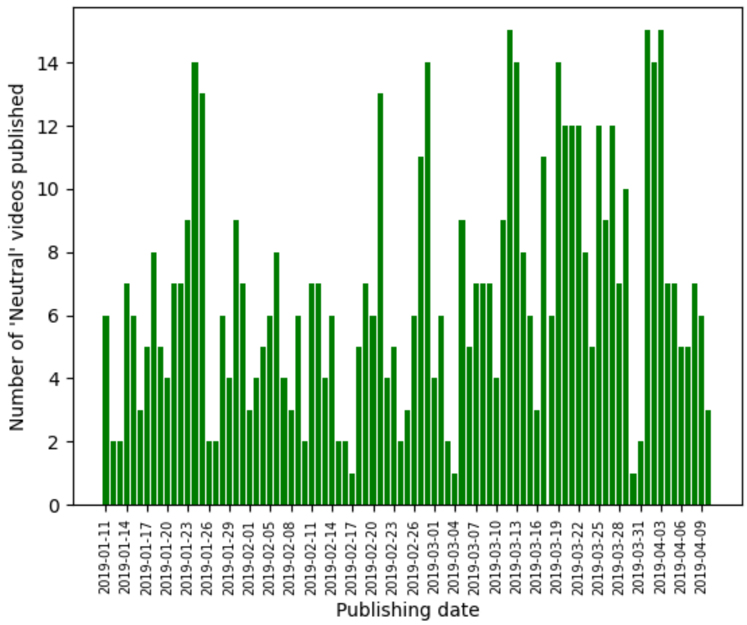

Figure 11: Temporal trend of Neutral videos (Indian General Elections) by all YouTube channels.

Figure 12 shows the temporal trend for ProGov videos. Similar to Neutral videos, we don't see a single peak in this graph, and the last peak sustains itself for an even extended time. The uploads stay high throughout the entire period and the dips are not as strong as the dips in Neutral and all videos.

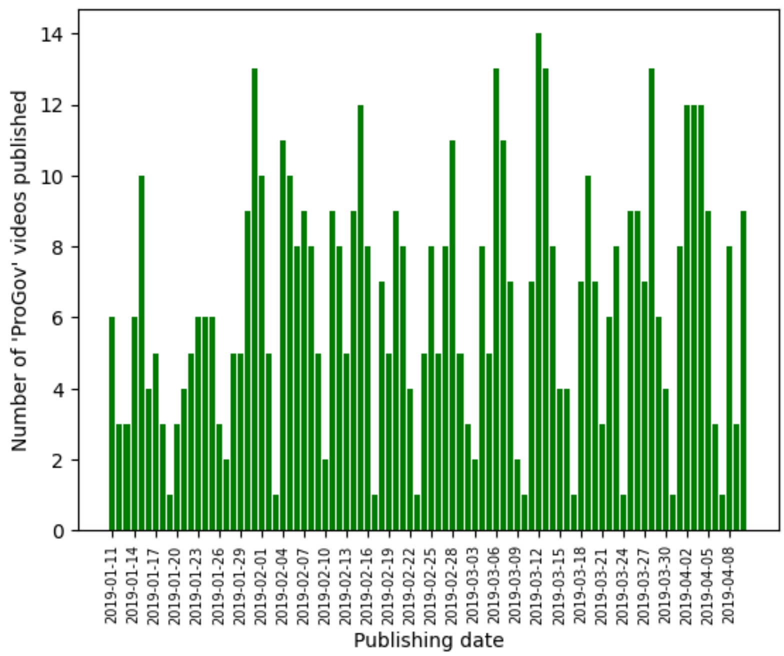

Figure 12: Temporal trend of ProGov videos (Indian General Elections) by all YouTube channels.

Figure 13 shows the temporal trend for AntiGov videos. We see a stark dissimilarity between AntiGov and all other types of videos. The number of AntiGov videos is always low. There is a global peak but it is very weak. Most of the days, the frequency of AntiGov videos drops down to one and stays there for the next couple of days. Moreover, the peak does not sustain itself and only lasts for a very brief period. Interestingly, the number of videos dropped to exactly 0 just a few days before the election dates, indicating that the news channel started pushing more ProGov and Neutral content as the elections approached.

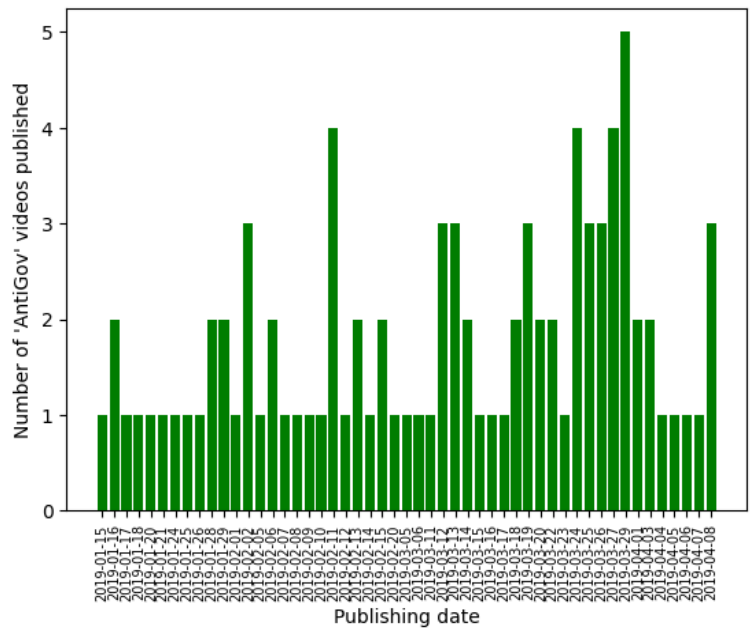

Figure 13: Temporal trend of AntiGov videos (Indian General Elections) by all YouTube channels.

Word clouds are graphical displays of text data in which the magnitude of individual words indicates their importance or frequency within the given text. A word will typically appear larger and more prominent in the word cloud the more frequently it occurs in the text. It is now simpler to quickly and easily identify the most important terms or themes in a document, webpage, or text body thanks to these visualisations, which also make it easier to quickly understand the main concepts or hot subjects.

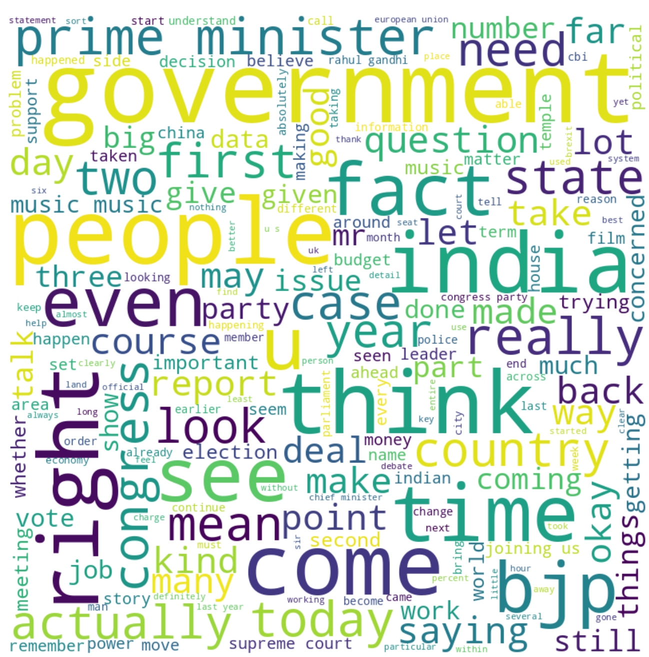

Figure 14: Word cloud of all videos (Indian General Elections) uploaded in January.

Figure 15: Word cloud of all videos (Indian General Elections) uploaded in February.

In January (figure 14) we can see that more national topics were in discussion such as national politics and national parties. But as soon as we come in February (figure 15), national topics are suppressed by debates related to 'Pakistan' and what are the prevalent conditions there. This gives us a very clear indication that the media no longer focuses on national development, whereas they discuss topics irrelevant to the Indian context.

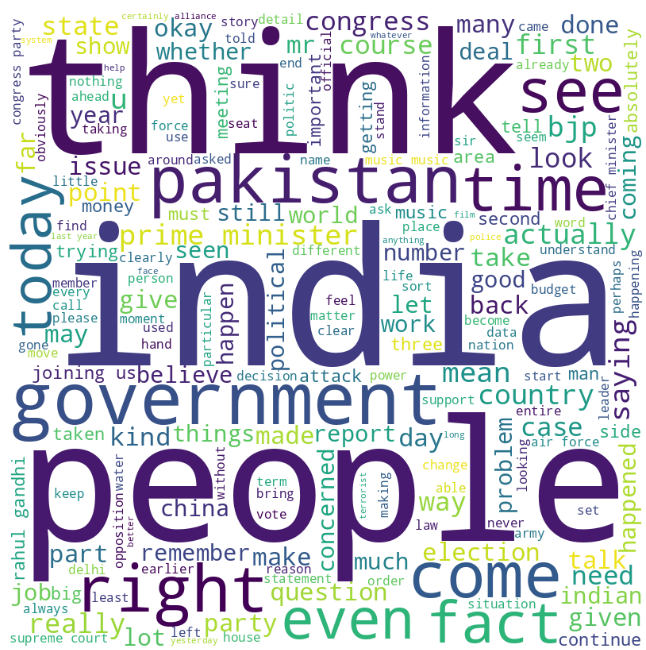

Figure 16: Word cloud of all videos (Indian General Elections) uploaded in March.

Figure 17: Word cloud of all videos (Indian General Elections) uploaded in April.

As we move ahead to March (figure 16), we can see no change in the agenda of media, as the prevalent topics in the media remain the same i.e. discussing the prevalent conditions in Pakistan. However, as we come closer to April (figure 17), which is the election month, Pakistan remains the hot topic, but a bit of national politics also comes into the picture. But the political topics are only centred on the general elections. Development and problems faced by the people remain missing from the mainstream media.

Karnataka State Elections 2023

Similar to the section on Indian General Elections above, Figure 18 qualitatively shows the type of content uploaded by specific YouTube channels before the Karnataka State Elections of 2023. The figure shows that the amount of Neutral and ProGov content on YouTube was much more than the AntiGov content.

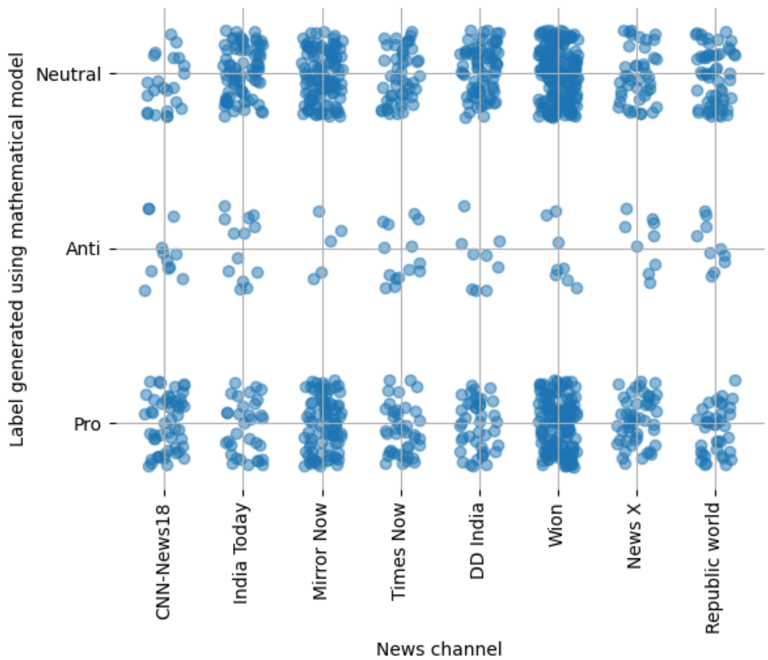

Figure 18: Qualitative distribution of videos (Karnataka State Elections) per YouTube channel.

Figure 19 gives a more quantitative version of figure 18. The figure shows that the number of AntiGov videos uploaded by each channel was less than 25, but at the same time, the number of ProGov and Neutral videos touched as high as 175. Moreover, we can see that the number of Neutral videos uploaded by India Today, DD India, Times Now, and Republic World were greater than the number of ProGov videos uploaded by the same channels, as in the Indian General Elections. This time Wion and Mirror-Now uploaded more Neutral videos than the ProGov videos as well. On the other hand, the number of ProGov videos uploaded by CNN News 18 and NewsX was greater than the number of Neutral videos uploaded by the same channels, as in the Indian General Elections.

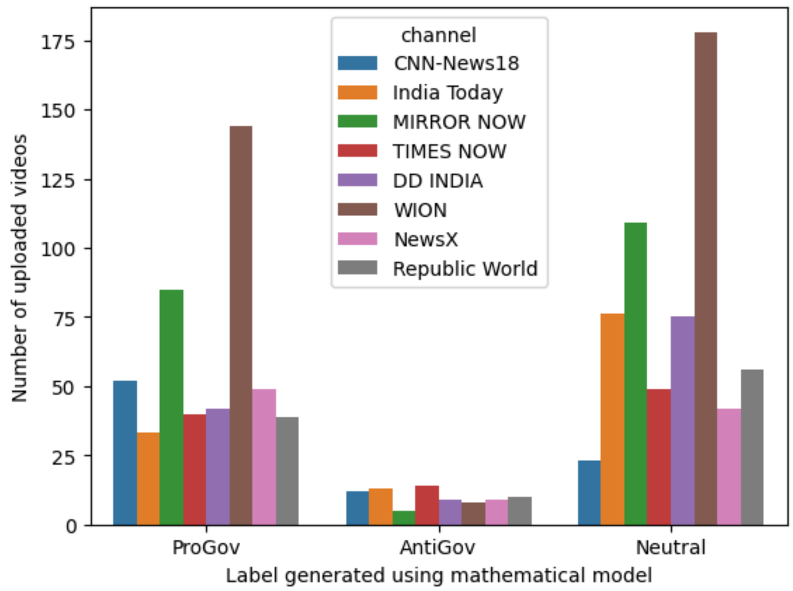

Figure 19: Quantitative distribution of videos (Karnataka State Elections) per YouTube channel.

Same as before, we now analyze the number of likes, comments, and views on the videos. Figure 20 shows the number of likes plotted as a violin chart against the type of video. We can infer that the number of likes of the ProGov and AntiGov videos is similar to what we saw for general elections. We can infer that the Neutral videos are gaining more likes than the ProGove and AntiGov videos. Analysing further, the number of likes in ProGov and AntiGov videos looks alike but the ProGov videos have a slightly higher number of likes. We must also consider that these are an absolute number of likes and the number of ProGov, AntiGov, and Neural videos are not the same in our dataset.

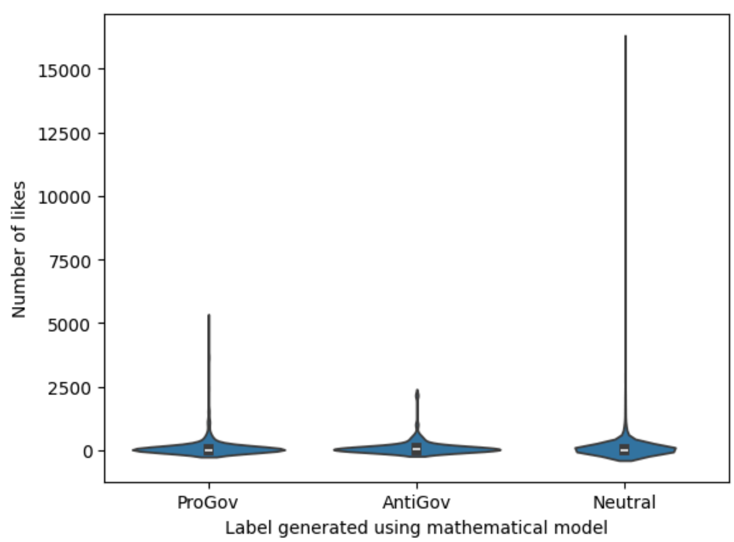

Figure 20: Numer of likes v/s type of uploaded videos (Karnataka State Elections).

Similarly, through figure 21, we can infer that the PrGov videos are gaining much more traction in terms of comments than the AntiGov and Neutral videos. The plot for the Neutral videos is extremely flat at 0, showing a scarcity of comments.

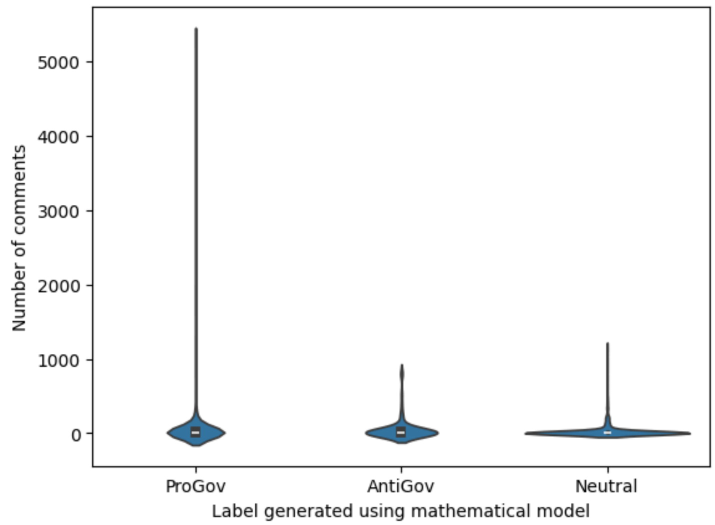

Figure 21: Numer of comments v/s type of videos (Karnataka State Elections).

In terms of views, figure 22 shows that every type of video seems to be gaining an equal distribution of views. The number of views is slightly higher for the ProGov videos and the lowest for Neutral videos.

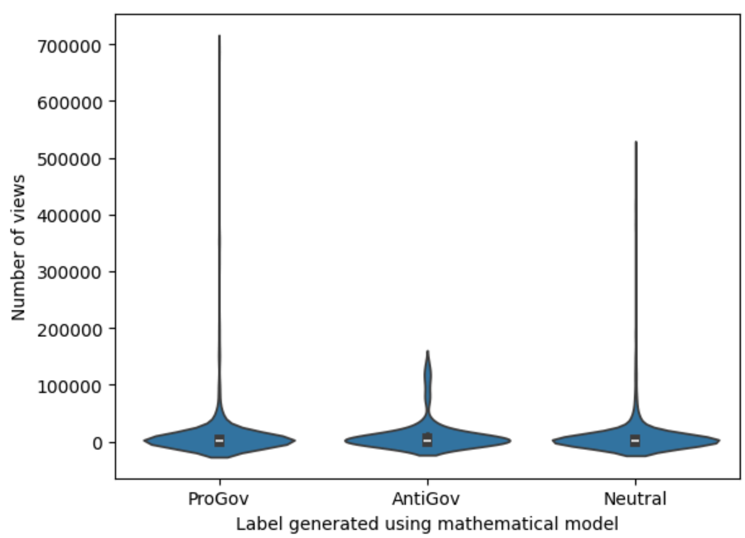

Figure 22: Numer of views v/s type of (Karnataka State Elections).

Lastly, we see the temporal trend in the number of different types of videos uploaded collectively by all the channels. Figure 23 shows a marginal trend of the number of uploads increasing before the elections and then decreasing back to normal as the election date approaches. More specifically, we see a peach around 1 month before the election date. There are also days when the frequency of uploads is comparatively low, even during the peak period. This plot is very similar to the plot created for the Indian General Elections.

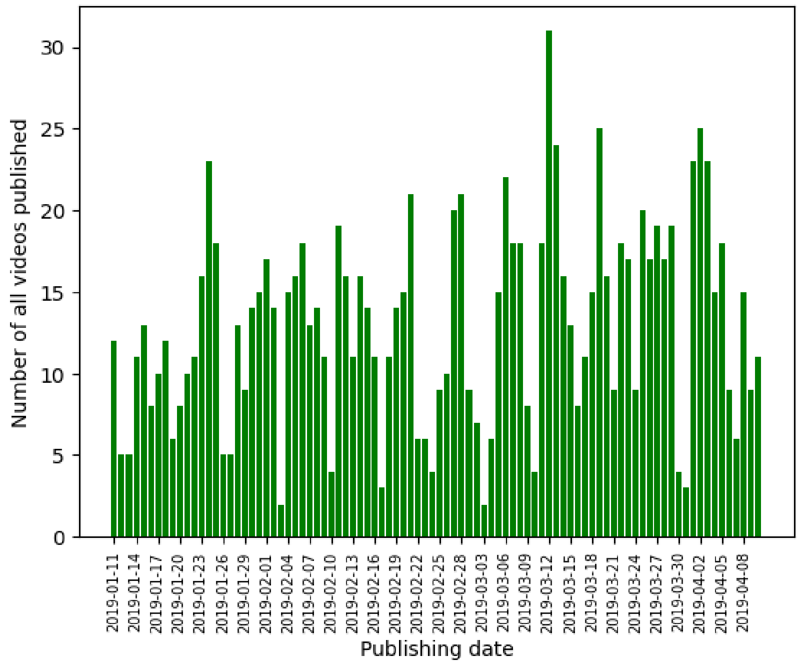

Figure 23: Temporal trend of all videos (Karnataka State Elections) by all YouTube channels.

Figure 24 shows a similar trend but only for the Neutral videos. We can see that this plot is more evenly spread out throughout the three months. Additionally, we see multiple peaks in this graph, one of which is one month before the election date. The Neutral uploads stay at an all-time high even after the total number of videos starts declining after the one-month peak.

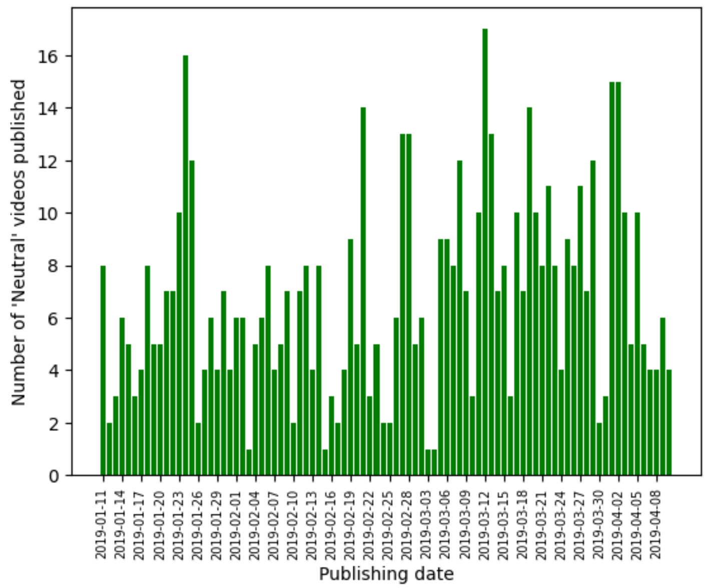

Figure 24: Temporal trend of Neutral videos (Karnataka State Elections) by all YouTube channels.

Figure 25 shows the temporal trend for ProGov videos. Similar to Neutral videos, we don't see a single peak in this graph, and the last peak sustains itself for an even extended time. The uploads stay high throughout the entire period and the dips are not as strong as the dips in Neutral and all videos.

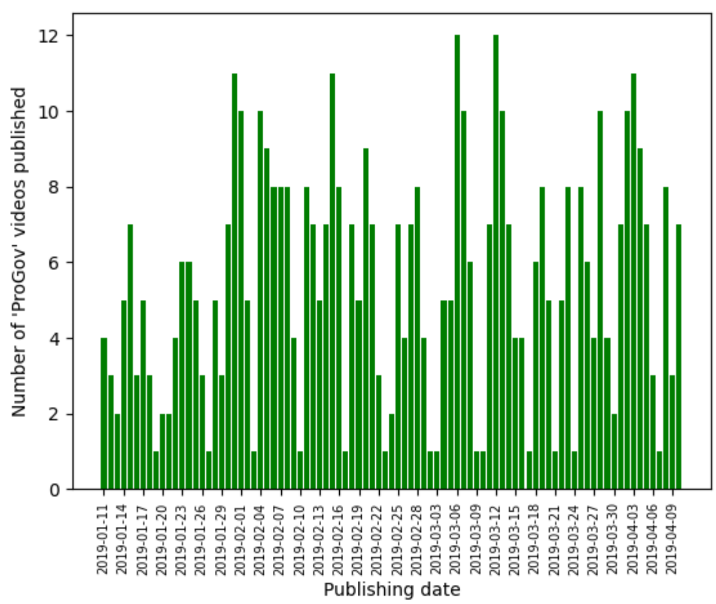

Figure 25: Temporal trend of ProGov videos (Karnataka State Elections) by all YouTube channels.

Figure 13 shows the temporal trend for AntiGov videos. We see a stark dissimilarity between AntiGov and all other types of videos. The frequency at which AntiGov videos are uploaded is always low. There is a global peak but it is very weak. Most of the days, the frequency of AntiGov videos drops down to one and the trend continues for a day or two. The peak also lasts for a very short period, not sustaining itself. Interestingly, the number of videos dropped to exactly 0 just a few days before the election dates, indicating that the news channel started pushing more ProGov and Neutral content as the elections approached.

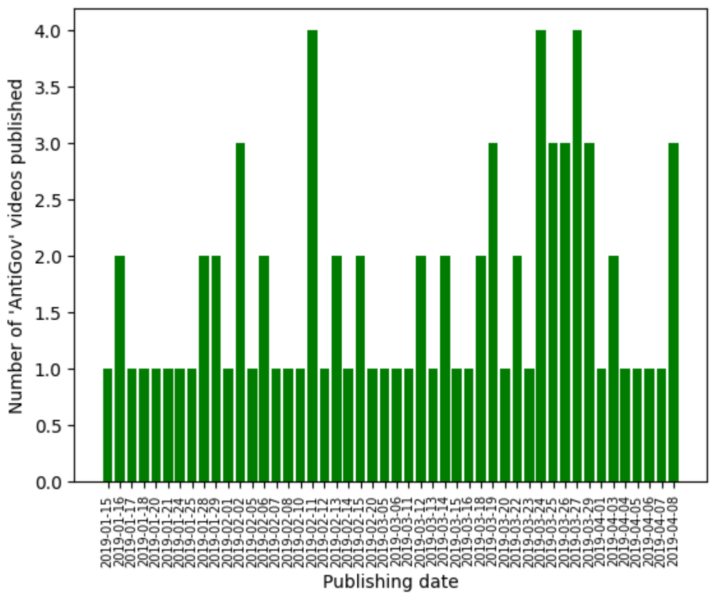

Figure 26: Temporal trend of AntiGov videos (Karnataka State Elections) by all YouTube channels.

Figure 27: Word cloud of all videos (Karnataka State Elections) uploaded in January.

Figure 28: Word cloud of all videos (Karnataka State Elections) uploaded in February.

During January (figure 27), hot topics in the news were relevant enough to national and state politics. But, as we head towards February (figure 28), we see that irrelevant topics to state elections become hot topics in the media channels. Concordant to what we saw in the general elections, Pakistan has yet again made its place in Indian politics hot topics, subsiding the development and topics of state interests.

Figure 29: Word cloud of all videos (Karnataka State Elections) uploaded in March.

Figure 30: Word cloud of all videos (Karnataka State Elections) uploaded in April.

The patterns are quite similar to what we saw in the general elections, Pakistan and irrelevant topics have been maintained in the hot topics, completely overshadowing national and state politics. From the word cloud, we observe that state political parties have been struggling to make it to the hot topics in March (figure 29). The same trend follows for April (figure 30), which was the election month in Karnataka.

Experimental Setup

The training and testing of the Deep Learning models were performed on a machine with 8GB RAM, Intel-i5-10300H CPU @ 2.5 GHz, and a dedicated NVIDIA GTX 1650ti GPU. The training of the Large Language Model was done via ChatPDF on the internet. Finally, Google Colab's environments were utilised to prepare, pre-process, and label the dataset via ou statistical model.

Results and Discussion

The statistical model used for preliminary classification does tell us that the most viewed English news channels in India post videos that support or favour the government rather than questioning the government or showcasing the opposition parties. Because the media does not hold the government responsible or tell the people about policies that reinforce the power of the incumbent, they may worsen autocratic behaviours and lead to democratic backsliding.

Viewership statistics of videos can be seen from the engagement people show towards the content. The basic stats of a video are its views, number of likes and number of comments. The viewership statistics of the videos are given below.

	ProGov Videos	**Neutral Videos**	**AntiGov Videos**
Total Videos	598	576	90
Total Views	6745561	6759262	1126504
Total Likes	81925	77734	10949
Total Comments	28314	16265	4656
Average views per video	11280	11734	12516
Average likes per video	137	134	122

Table 10: Viewership summary and engagement stats for General Elections 2019

	ProGov Videos	**Neutral Videos**	**AntiGov Videos**
Total Videos	484	608	80
Total Views	5575633	6655261	966788
Total Likes	66829.0	78067	9010
Average views per video	11520	10946	12084
Average likes per video	138	128	112

Table 11: Viewership summary and engagement stats for State Elections 2019

From the above viewership stats, we infer that despite having a higher average view per video, people tend to appreciate the ProGov videos more than AntiGov videos. This also concurs with the fact that the ruling party did win the election in the General election but ended up losing the Karnataka election.

Once the data has been classified, we use advanced natural language processing (NLP) models, BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly Optimised BERT Pretraining Approach), BART (Bidirectional Auto-Regressive Transformers), and Distilled BERT to detect bias in video transcripts. The models were trained over an 80:20 train-test split of the dataset. The evaluation metrics of the models are shown below in tables 12 and 13. The tables show that BART performs the best on our dataset.

Model name	**Training accuracy**	**Testing accuracy**
BERT	91.51	56.84
RoBERTa	99.9	73.12
BART	**99.9**	**74.7**
Distilled BERT	99.4	69.56

Table 12: Evaluation of trained models for Indian General Elections.

Model name	**Training accuracy**	**Testing accuracy**
BERT	91.03	60.85
RoBERTa	99.89	76.17
BART	**99.89**	**77.02**
Distilled BERT	99.67	67.65

Table 13: Evaluation of trained models for Karnataka State Elections.

We further discuss the answers of the Large Language Model we trained over the subset of our data. The questions and answers are listed in table 11.

Question	**Answer**
From the text,	From the text, it can be inferred that the government of India has

what can you infer about the government of India?	been involved in a contentious legal battle regarding the Raphael papers, with objections raised about the admissibility of the documents. There are discussions about the government's attempts to resist the probe into the acquisition of Raphael jets, citing national security reasons and alleging that the documents were stolen from the defence ministry. Additionally, there are references to allegations of corruption and questions about the government's transparency and accountability in the Raphael deal. The text also highlights the government's interactions with the Official Secrets Act and the implications of the Supreme Court's ruling on the government's actions and motives. Overall, the inferences about the government of India from the text suggest a complex and controversial situation surrounding the Raphael deal and the government's response to it.
Apart from the Raphael deal, what can you infer about the government in general?	The text does not provide a comprehensive analysis of the government in general, as it primarily focuses on the legal battle surrounding the Raphael deal. However, there are some references to the government's handling of issues related to freedom of the press and the right to information, with discussions about the Official Secrets Act and the Supreme Court's decision on the admissibility of the Raphael papers. Additionally, there are mentions of political rhetoric and campaigning related to the upcoming elections, with comments from representatives of both the BJP and Congress parties. Overall, while the text does not provide a detailed analysis of the government in general, it does suggest that there are ongoing debates and controversies surrounding issues related to transparency, accountability, and freedom of expression in India
Should I vote for the current government based on this text?	Based on the text provided, it primarily focuses on the legal battle surrounding the Raphael deal and related political discussions. The text does not offer a comprehensive analysis of the overall performance of the government across all areas of governance. Therefore, it may not provide sufficient information to make a fully informed decision about voting for the current government. It is advisable to consider a wide range of sources and factors when making decisions about voting, including the government's performance in various domains, its policies, and its alignment with individual values and priorities.
Is the text pro-government or anti-government?	The text contains a variety of viewpoints and statements from different individuals, including representatives from both the BJP and Congress parties. It discusses legal proceedings related to the Raphael deal and includes comments on issues such as freedom of the press, the Official Secrets Act, and the Supreme Court's decisions. The text

	does not appear to have a clear overall bias in favour of or against the government. Instead, it presents a range of perspectives and discussions related to the legal and political aspects of the Raphael deal and its implications.
What do you infer about BJP as a political party from the text?	The text provides information about the BJP's stance on the Shabarimala issue and its manifesto related to faith, tradition, and worship. It also includes comments from a BJP spokesperson and discusses the party's efforts to benefit electorally from the anger surrounding the Shabarimala issue. However, the text does not provide a comprehensive analysis of the BJP's overall performance or policies across all areas of governance. Therefore, it may not be possible to make a definitive inference about the BJP as a political party based solely on the information contained in the text.
What do you infer about the opposition party from this text?	Based on the text provided, it appears that the opposition party, particularly the Congress, has been involved in discussions and negotiations regarding alliances with other political parties in various states. The text mentions the Congress party's efforts to stage alliances across the country and its attempts to build a narrative focused on saving democracy. Additionally, there are references to the Congress party's decision-making process regarding alliances and its positioning in certain states. However, the text does not offer a comprehensive analysis of the overall performance or policies of the opposition party. Therefore, it may not provide a complete basis for making a definitive inference about the opposition party as a whole.

The answers reveal that the model does not favour a particular political party or ideology. It does emphasize the on the current discussions of freedom of speech going around in the country, but at the same time it talks about the media acknowledging the discussion. The answers talk about the shortcomings of the current government, which means that the media is discussing these points on YouTube, but the the number of videos in which they discuss these points are way too less when compared with pro-government videos. However, given the limit of the textual information allowed as dataset in the free version of ChatPDF, the model cannot form a rigid opinion of any political party. The answers do put BJP in a negative image and Congress in a positive image, but it is important to note that this is only a subset of the data, and the model does not form any opinions after processing the complete text.

Conclusion and Future Work

Through this study, we can conclude with high confidence that the news media tend to upload more ProGov and Neutral videos on their YouTube channels near the election date. The amount of AntiGov videos stays at an all-time low and completely disappears a few weeks before the elections. Furthermore, consumers interact more with ProGov and Neutral videos than they do with AntiGov videos. The interaction of people with only ProGov and neutral news makes them live in an Eutopian bubble, where they assume nothing is wrong. From the word clouds created for the pre-election period, it becomes very clear that the media is negligent towards national and state matters. The main focus of the media remains talking about Pakistan and irrelevant matters, only highlighting the plus points of the government, instead of questioning the government over topics of public interest.

The pre-trained textual models perform decently on the collected and mathematically labelled dataset, with the highest training accuracy of 99.9% and the highest testing accuracy of 74.7% for Indian General Elections. Similarly, for the Karnataka State Elections, the model achieved a training accuracy of 99.9% and a testing accuracy of 77.02%. Finally, the Large Language Model trained on a subset of the dataset performs neutrally on the asked questions. The model does not favour a particular party, but that can also be because of the security mechanisms built in the language model.

This work can be extended to include more elections from the Indian history, especially the elections where the working government changed, to get a more holistic answer to our questions. Additionally, more tectual models can be trained to achieve a better testing accuracy on our dataset. Finally, with enough resources, a custom Large Language Model with no security barriers can be trained to give a transparent view of the type of the content uploaded on YouTube near Indian elections.

References

1. Bessi A, Zollo F, Del Vicario M, Puliga M, Scala A, Caldarelli G, Uzzi B, Quattrociocchi W. Users Polarization on Facebook and Youtube. PLoS One. 2016 Aug 23;11(8):e0159641. doi: 10.1371/journal.pone.0159641. PMID: 27551783; PMCID: PMC4994967.
2. Emily Kubin & Christian von Sikorski (2021) The role of (social) media in political polarization: a systematic review, Annals of the International Communication Association, 45:3, 188-206, DOI: 10.1080/23808985.2021.1976070
3. Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). Examining the consumption of radical content on YouTube. Proceedings of the National Academy of Sciences, 118(32), e2101967118. https://doi.org/10.1073/pnas.2101967118
4. https://eci.gov.in/statistical-report/statistical-reports/
5. Eberl, J., Boomgaarden, H. G., & Wagner, M. (2015). One bias fits all? Three types of media bias and their effects on party preferences. Communication Research, 44(8), 1125–1148. https://doi.org/10.1177/0093650215614364
6. Eberl, J., Wagner, M., & Boomgaarden, H. G. (2016). Are perceptions of candidate traits shaped by the media? the effects of three types of media bias. The International Journal of Press/Politics, 22(1), 111–132. https://doi.org/10.1177/1940161216674651
7. Lin, Y., Bagrow, J. P., & Lazer, D. (2011). More Voices than Ever? Quantifying media bias in networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1111.1227
8. Dasgupta, B. (2022). Indian media: How bias and fake news originate from the country's politics. Ijpmonline, 1(1), 45–48. https://doi.org/10.26524/ijpm.1.7
9. Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805.
10. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, & Veselin Stoyanov. (2020). Ro\BERT\a: A Robustly Optimized \BERT\ Pretraining Approach.
11. ChatPDF - Chat with any PDF! (n.d.). https://www.chatpdf.com/
12. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, & Luke Zettlemoyer. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.

13. Victor Sanh, Lysandre Debut, Julien Chaumond, & Thomas Wolf (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR, abs/1910.01108.