Analysis of Tweets related to Indian elections-2019

Table of Contents

1. Introduction

Sentiment analysis is one of the largely used text mining techniques in data science. It is at its best in analysis of customer reviews and feedbacks on a particular product. It is widely used across domains like travel destinations, hotel bookings, products listed on e-shopping websites etc. In this visualization project Firstly, I would like to show the politicians (Narendra Modi and Rahul Gandhi) as what people are speaking about them. Through this analysis, the politicians can improve themselves and put more emphasis on the needs which will eventually develop the country.

Secondly, the most awaited election results where the whole country would be eager to know about it. Hence, I would like to show the prediction results of each state to citizens of India through sentiment of the people in each state.

Thirdly, Election commission of India has released the dataset of voter turnouts in each state which would provide numbers about total male and female electors and people who actually turned up for voting. The intended audience would be again government of India. The analysis would help to identify the particular state which has low percentage and can dig deeper to determine the root cause for the issue. Lastly, I have discussed few comparisons between Narendra Modi and Rahul Gandhi to determine the probable prime minister of India.

2. Design

The given problem statement is to analyze-2019 Lok Sabha elections of India and determine most probable Prime Minister of India and winner of each state through twitter sentiment analysis. To successfully perform this task the major the requirement is the collection of the data from twitter and cleaning which has been completed in the previous exploration project in R and then visualizing through D3. The main ideas are as follows

Ideas:

Method/Process to determine the problem statement

- ✓ The prediction can be done through comparing historic elections wins in each State and plotting the same on different kinds of charts.

- ✓ Analyzing the current prime ministers/chief ministers achievements and other influencing parameters and visualizing it.
- ✓ Analyzing the election turnouts percentages in each state and determining the some statistics about Male and female ratio and reasons behind the same. Displaying the results through multiple charts.
- ✓ Using Sentiments of each state to determine the each winner and also same sentiments can be used to determine the Prime Ministers.

Graphics/Charts:

- ✓ Pie chart to compare different political parties in each state that is basically to check number of time a political party has come to power.
- ✓ Circular Bar plot to show turnout percentages.
- ✓ Bar plot to count number of achievements/bad marks of each political party
- ✓ Categorical Choropleth map to show winner of each party.
- ✓ Sankey diagram to show some text analysis
- ✓ Grouped bar between Modi and Rahul's sentiment scores
- ✓ Word cloud of both Modi and Rahul.
- ✓ Gauge chart to determine the popularity.
- ✓ Scatter plot to show distribution of tweets.
- ✓ Grouped Bar chart to show total elector and total voters in each state.
- ✓ Line plot to show time V/s number of tweets related to politician.
- ✓ A world map to show tweet distribution of Modi and Rahul
- ✓ Bar Plot to show distinct languages used to tweet.
- ✓ Pie chart to compare Voter Men and Female ratio.
- ✓ Bar graph to shows sentiments score towards each wining state.

These are the initial ideas generated through brainstorming. Out of which I try to remove some of the methods which wouldn't give better analysis results and also some charts which portray same ideas in a different way.

Initial idea which said to predict the winners by means of checking historical records would be redundant as situation change every year and election results are more related to contestant than the political party. Secondly, analyzing the current prime ministers achievements/bad marks would be favorable option as people are more concerned about the current scenario than the history. This would lead to analyze sentiments of people with respect to each contestant or state to determine the winner. Furthermore, analyzing turnout percentage would give more insights are why citizens of India are turning back towards the elections. Therefore

the last two points on method would be in line with the problem statements and hence I have chosen this idea.

The only disadvantage of this would be collecting the data from past. From Developer API twitter version only provides a week data. The volume of data is huge and computational time to clean up and getting sentiments score would be high.

With regards to the selection of the appropriate charts and positioning of it on the webpage I have chosen some of the charts from the above lists like pie chart to show differences between male and female number of voters, bar graph showing sentiment score of each state, Choropleth showing Voter turnout percentage and categorical map displaying the winning political party in each state.

To compare the differences between opinions of the people about the two leaders I have chosen gauge chart for displaying popularity between Modi and Rahul, word clouds between them with respect to elections, grouped bar chart which shows different kinds of sentiments like happy, joy, disgust, negative, positive etc between the two politicians, a world map which displays the spread of both leaders across the globe.

I could even consider sankey chart to show word connection between most frequent words which would be unnecessary for analysis. Circular bar plot is not that communicative for the turnout percentages as most of the values are on the same line. To show time series and trend in the data about the people's sentiment we don't have enough dates. Tweet distribution on the whole is not that important in the analysis and hence discarding it.

The final product is divided into 3 sheets where each sheet conveys a particular component of the problem.

Sheet 1: The main aim of the project is to analyze the sentiments of the each state and predict the wining party. Hence this sheet would have a categorical Choropleth with the legend at the bottom which shows political party and total number contestants won (it is called number of seats won for a particular party in terms of political language). Just to give add on to the webpage, pictures of two probable prime ministers are added along the gauge chart showing popularity. A tool tip is defined to on the pictures to show some content about the pictures. Furthermore, two buttons are created at the top of the page so as to link between the other sheets.

Sheet 2: It consists of comparison of two popular and probable prime ministers of India through sentiments obtained. Initially word clouds shows the difference of opinions among the people and added link to each word which automatically redirects on-click to Google page showing details word associated to each one of them. Grouped Bar chart showing sentiment comparison

between the two. Two KPI objects which display languages used to tweet and how far words "Modi" and "Rahul" have reached across the globe. As usual buttons to link between the sheets

Sheet 3: It basically aims to convey information about the turnout percentage in each state. However, along with that information it also shows bar which displays sentiment scores of the winning political party. The page layout is designed with Choropleth of turnout percentage at the center, a pie chart at the top left showing Male and Female voter ratio and bar chart of sentiment score at right bottom. The link is created to election commission of India website along with the button to link other sheets.

Data Sources:

There are 29 States and 7 Union territories in India, thus whole dataset has roughly 90k-100k Tweets. Data is collected in such a way that for each state we take ruling party and opposition party has tags and tweets equally so as to maintain zero bias. Voter turnout dataset is taken from https://eci.gov.in/poll-turnout/

Updated Final design Sheet

As the saying goes "A picture is worth of thousand words"– I tried to make some corrections in my final design to impact more on human perceptual system.

When I started implementation, I thought of merging the sheet1 and sheet2 because both sheets are having some kind of comparisons like sheet 1 has gauge chart which is showing comparison between the two politicians and sheet2 has word clouds. Firstly thinking in terms human perceptual system having the two sheets of together would more impact towards human system as both convey differences.

So when I merged it was too much packed with multiple graphs and hence I have removed grouped bar chart which was just showing frequency of positive and negative sentiments which are anyway portrayed through word clouds. These word clouds has more impact on brain as it has frequency of word is shown by size of the word. Instead of grouped bar which was consuming lot of space I added a globe which shows the spread of word Modi across the globe. Though it takes time to count distinct countries, it has good visual effect. The other charts like categorical and density Choropleth, Pie charts are visually interpreting as they have impressing colors. Lastly I added pictures to website to add more impact because when the user looks at it he definitely understands that there is some type of comparisons being carried out through analysis. Rest all KPIs objects remain same.

Hardware/ Software/ Time/Resource requirements

Ideally it should work on any system with Mozilla Firefox browser. There are chances due to pixel issues the alignment may change. Time spent on this project is around 1 week (approx. 50 hours). I under estimated the budget and also other subject commitments, it would be much better if I had given myself 2 weeks of time to implement as I am new to environment.

3. Implementation Reasons

The Motivation behind the project is development of the country. Politics being one of the major topics where citizens of India will more interested in discussing and knowing the facts. Election which is currently happening is one of the hot topics on internet. Analyzing people's emotions towards it will help all the politicians know their positives, negatives and also improvements required in the society.

Environment/libraries used

Majorly for visualization I have used D3 and Bootstrap library of JavaScript and R for data manipulation.

Why D3 not R?

D3 can handle large amount and has flexibility towards using HTML, CSS and SVG elements added to it. One can create their own visualization in d3 than using any packages in R.

Type of charts implemented and why?

Gauge Chart: It is used when there is an exact one point to tell on the given dataset. Hence in my case I have used to show the popularity of the politicians on tweets. My aim was to determine the upcoming prime minister and hence comparing the popularity would be one of the key points in predicting the prime ministers.

Categorical Choropleth map: These maps are great to show regional pattern in the data. For instance, it the best to show winner of each state in the election through different colors and to make it more sensible a legend is added. Therefore I chose this chart which would convey the audience about the winner of each state. Along with the legend I have added total points secured by the each political party which would be helpful to determine the overall winner.

Word Clouds: These identify trend and patterns available in the data. Frequent words when viewed in tabular format may be overlooked but size of the word determines frequency of the word in word clouds and hence it is easy to determine the pattern. I am using this because I want to convey the politicians as what people are speaking about them which would indirectly help in development of the country if actioned upon it. Click feature on the words are added to it so when clicked it redirects to webpage which would show content related to the word with the politician.

Globe: This is a bit fancy where the globe rotates highlighting each country at a time. I have used this to show the Modi's spread across the world. These are basically tweets recorded for Modi from different parts of the world irrespective of polarity of the statements. This basically compares how far Modi / Rahul have reached across the globe which would essentially determine popularity in different way.

KPI objects (text boxes): These are objects to show numbers which indicate a particular feature. In my case I have introduced this because aim is to compare probable prime ministers and inter the upcoming prime minister. These object shows different languages used to tweet for both of the politicians by which we can determine the popularity again.

It's is very clear question why I am using different methods to compare the popularity. This is because we are not checking the polarity other than that of the word clouds and hence it required to analyze through different methods to come to conclusion about the winning Prime Minister.

Choropleth Map: I have used this map to show to voter turnout percentage in each state. This would help to identify the state with less percentage and dig deeper to understand what the root cause for the issue is. The tool tip shows with winning political party along with total voters and electors. I have added winning party to visualize if there are correlations between winning state and turnout percentage.

Pie chart: Pie chart is helpful to understand the ratio of each quantity and it is good to use if there are not much categories. Therefore I am using this to show Male and female ratio of voters. It basically shows for every 100 male voters how many females have voted. This would show countries development through women empowerment.

Bar chart: I have used this to show sentiment score obtained by each state towards winning state. This would help understand the countries progress through digitization.

4. User guide

The folder contains html file named index.html. To view the webpage created, open the file using Mozilla Firefox. One of the problem statements which I wanted to convey through visualization was to show winner of each state. The figure below shows part of it.

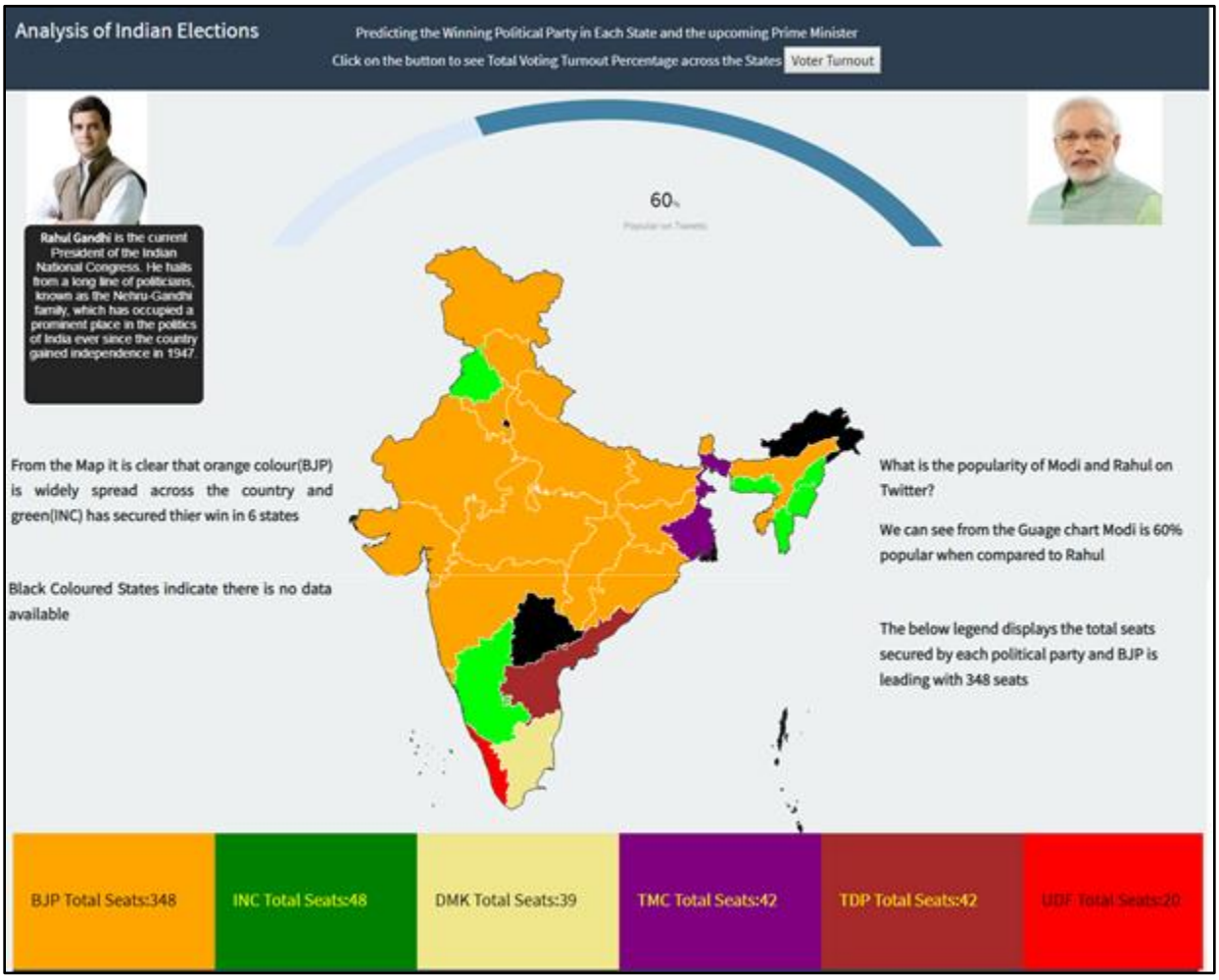


Figure 1: Shows the Political Party distribution throughout the country

I have provided an option to view some details about the two pictures on hover as show in figure (1). There is hyperlink on the header which will redirect the webpage to voting turnout percentage sheet which will be discussed in later stages. On the sides there are brief descriptions on the key observations.

There is no interactivity defined for map as it is at its best when shown without it. As an alternative a legend kind of boxes are added at the bottom which will show up total number of seats won for each political party.

Second goal of the visualization was to so show the comparison between the Narendra Modi and Rahul Gandhi which is achieved by using Word Clouds and also gauge chart as shown in the figure.

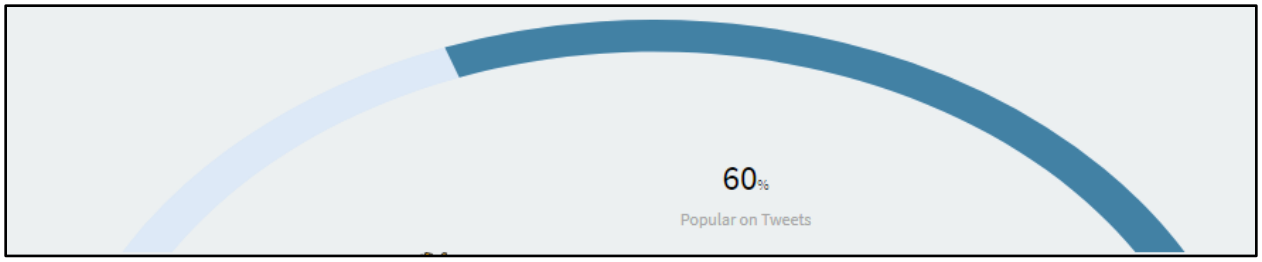


Figure (2) – Showing the popularity of Modi VS Rahul

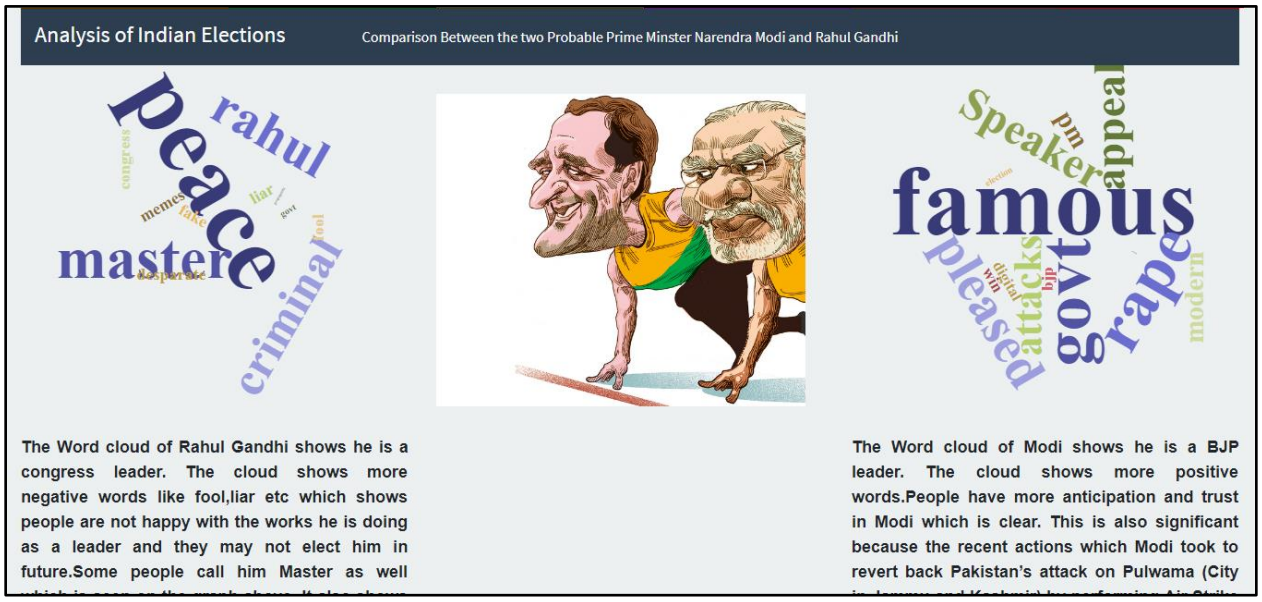


Figure (3) - Showing the word cloud of both the politician

Simple interactivity is achieved when clicked on words where it directs to Google page showing the contents related to the person with that word. For instance, if you click on word "government" on Modi's word cloud it searches on Google for Modi's government.

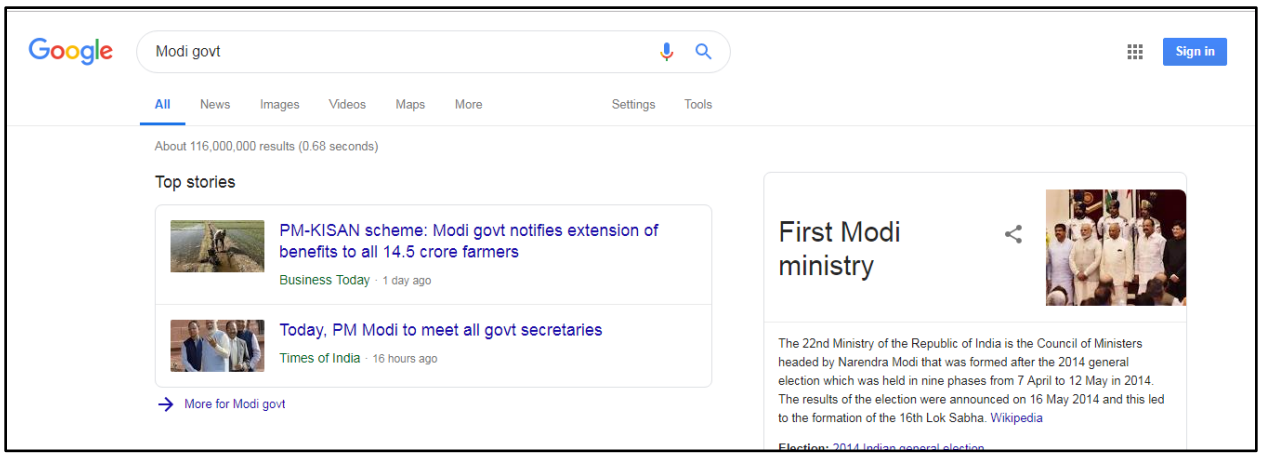


Figure (4) – Output of click on govt in word cloud

A picture in the middle of figure (3) is displayed to make the webpage more attractive which portrays the race between Modi and Rahul in the election.

There are two green colored buttons of Modi and Rahul when clicked on it displays distinct languages tweeted for the tweets extracted from them. Default number button is set to Modi's distinct languages count.

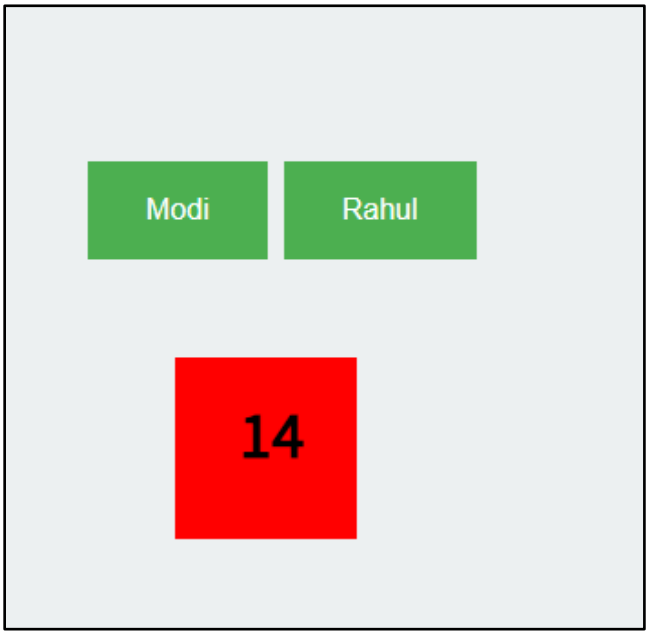


Figure (5): Showing the distinct languages used to tweet with buttons

It is the animated globe hence no interactivity is added. However it rotates highlighting the countries where Modi's has been tweeted from and displays name of the country below it.

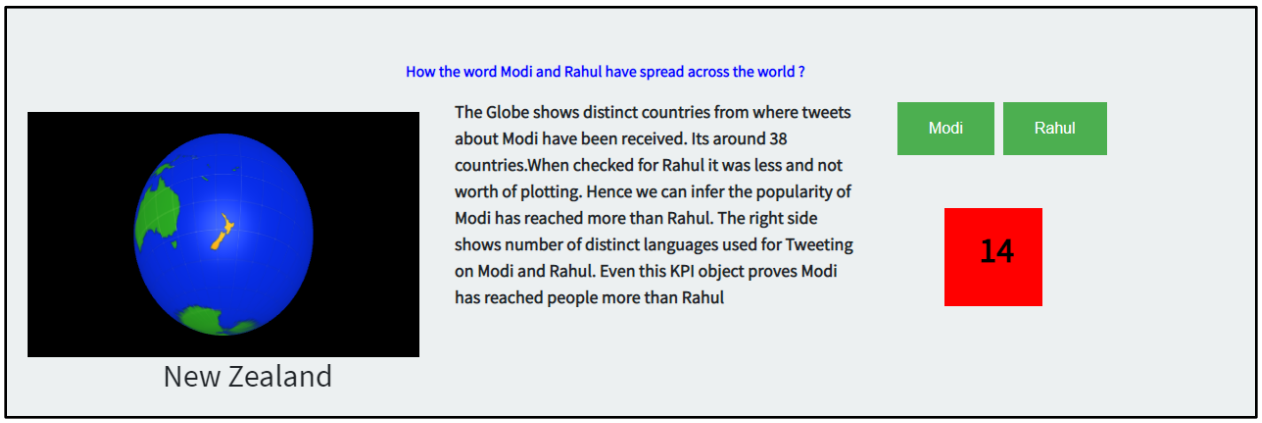


Figure (6): Showing the Globe which displays Modi's reach across world

As discussed initially when clicked on voter turnout, it redirects the sheet to Voter turnout page where an interactive Choropleth is found. Color intensity is dependent on the voting turnout percentage in each state. When hovered over the states it displays a message which shows turnout percentage, total voters and seats available and winning political party for every state. On the right hand side it pops ups pie chart and bar chart where pie chart represents for every 100 male voters-number of females electors who voted and bar chart represents the sentiments wining political party in each state. Both the graphs pop up on hover over Choropleth.

Lastly green button which says sentiment analysis when clicked, it takes to the index page and the header which says "Indian Election-2019" when clicked takes to the Election Commission of India site where data is downloaded for turnout percentages.

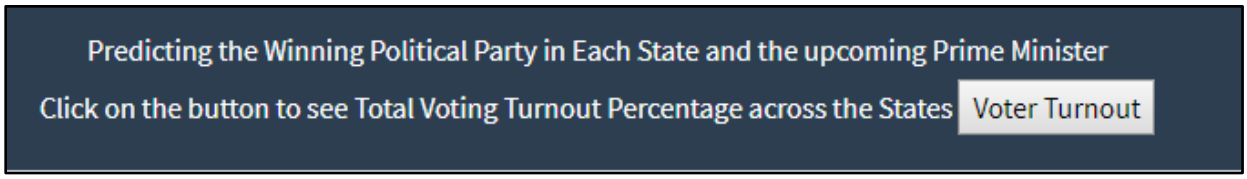


Figure (7): Button to Navigate

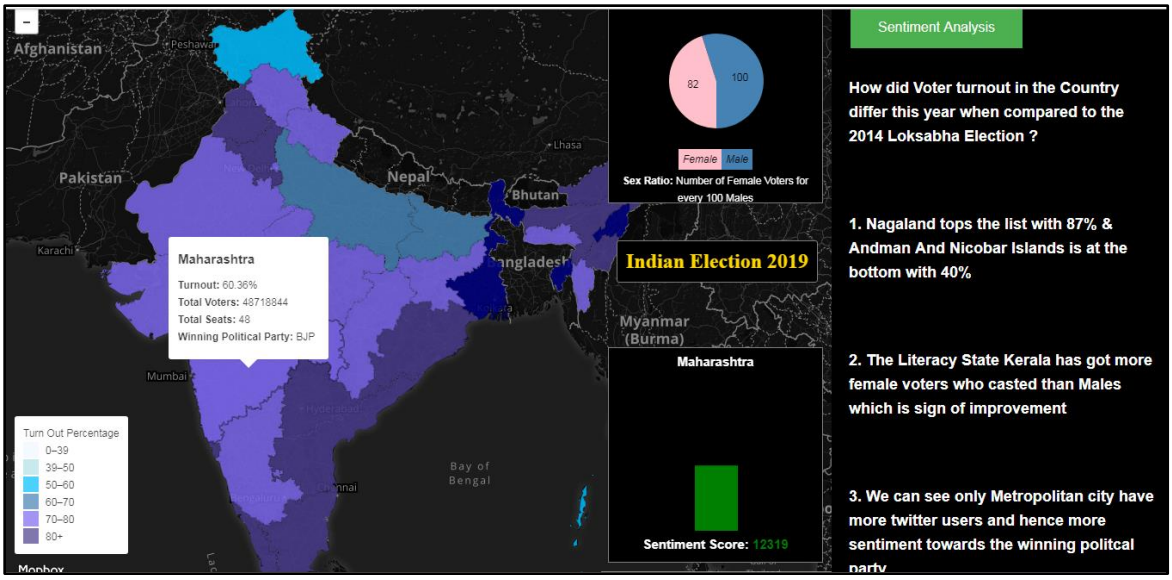


Figure (8): Choropleth of the Turnout percentage with Pie and Bar charts on hover with tooltip and button to navigate to index sheet

5. Conclusion

Completing my visualization project in D3 is one of the best achievements in this semester. The results of the election are released and my predictions about the winner in each state are nearly 70% accurate. Through the project I have learnt twitter data sentiment analysis and some different kind of visualization which can be applied to visualize the same. At the same time I have understood how hard it is to determine the sentiment of the sentences as some of the words can be positive or negative and also it hard to analyze the sarcastic tone of the sentences. On the other hand visualizing the same on D3 was tough task and I have learnt to implement visualization from scratch without using any libraries.

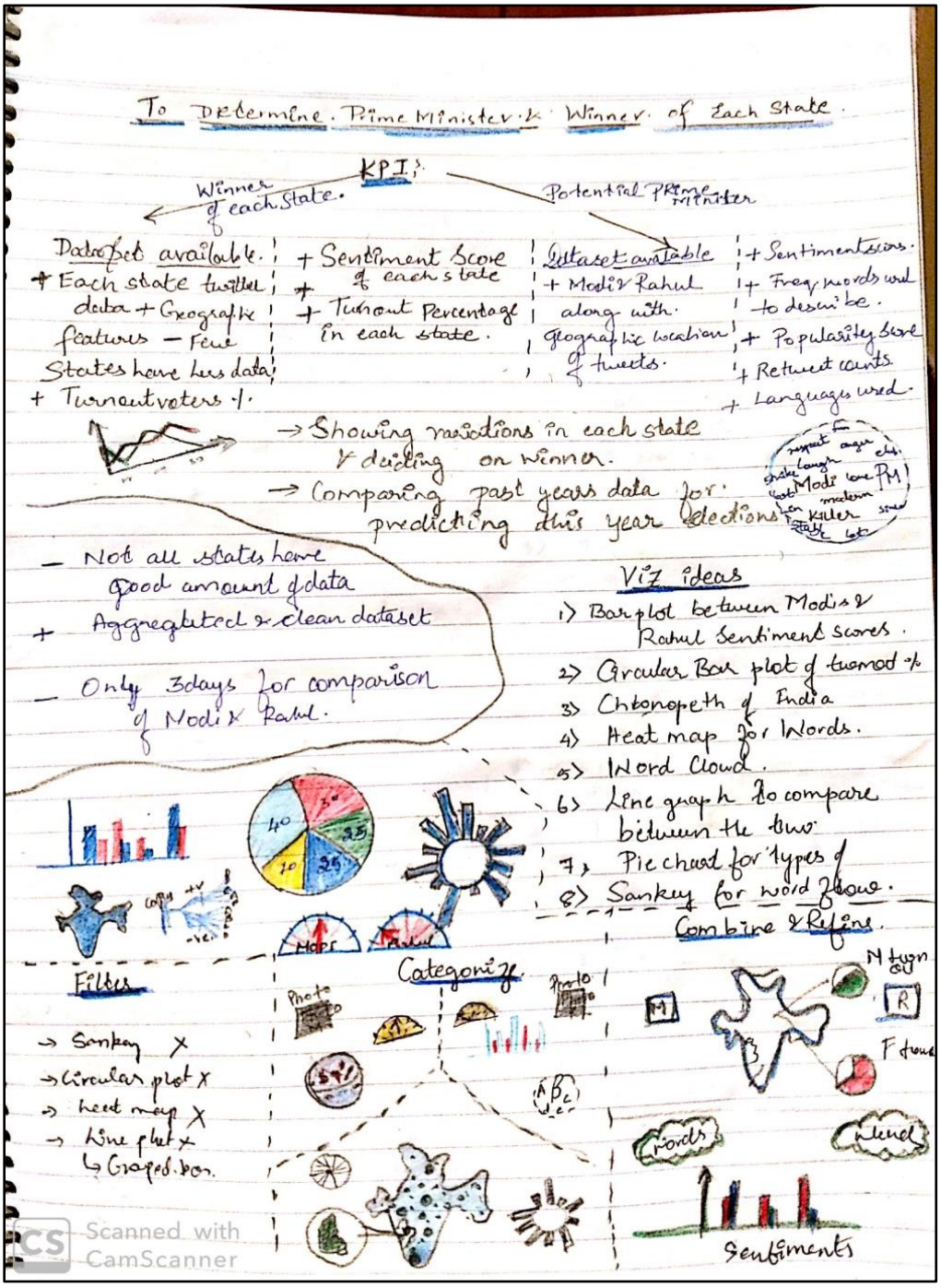
If given a chance to re-design / some extension in the time I would like to modify my code to fit any type of screens because I see some discrepancies in pixels when I try to put it into my friends system.

6. References

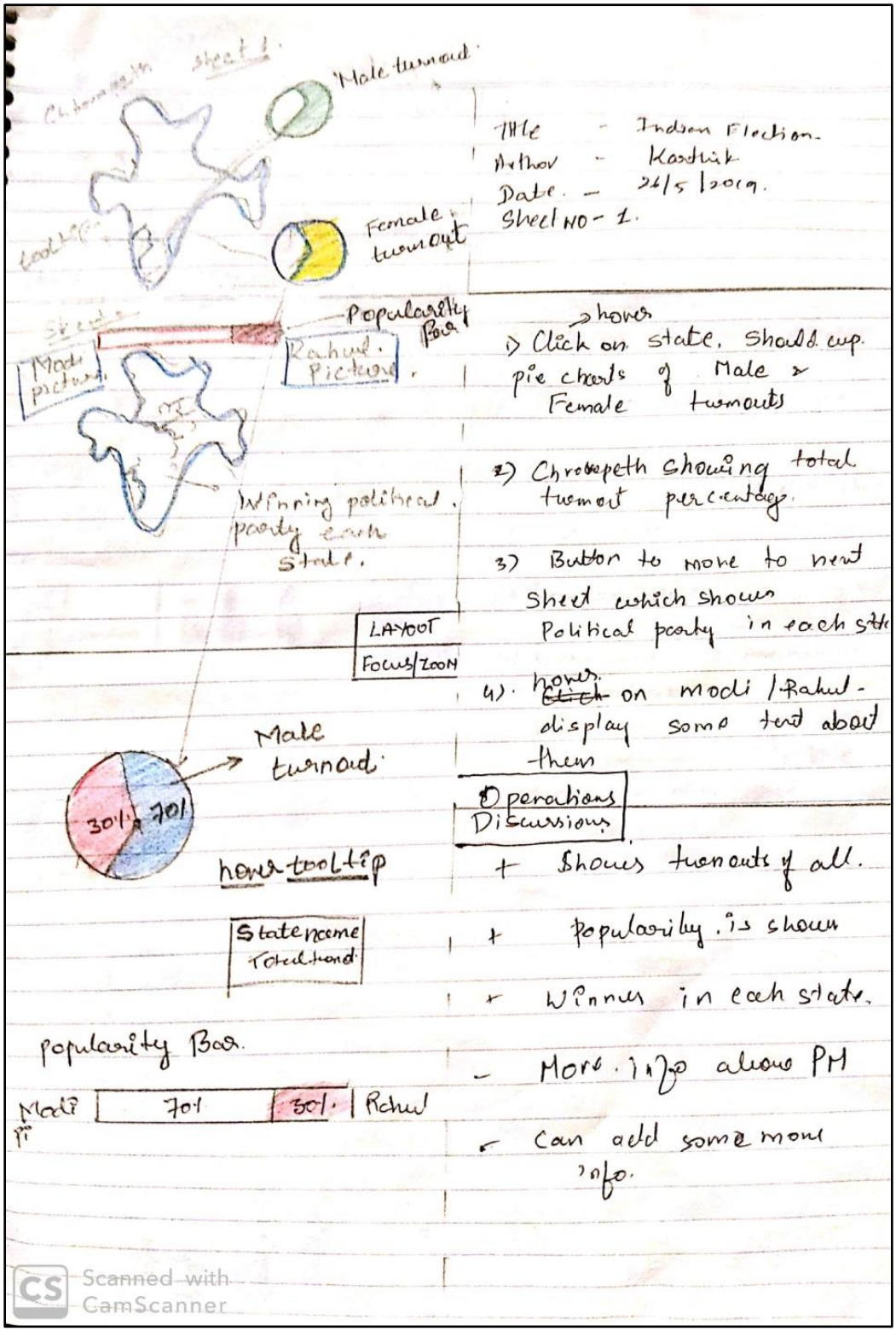
- https://www.d3-graph-gallery.com/pie
- https://www.d3-graph-gallery.com/graph/choropleth_basic.html
- https://observablehq.com/@d3/world-tour
- https://github.com/jasondavies/d3-cloud
- https://bl.ocks.org/allisonking/ece2f8a08a626b7067381317a385a245
- http://bl.ocks.org/msqr/3202712
- http://bl.ocks.org/JulienAssouline/209554f6002f2464e328495a14752830
- https://github.com/wvengen/d3-wordcloud
- https://github.com/shprink/d3js-wordcloud
- https://github.com/jasondavies/d3-cloud

7. Appendix-1 Design Sheets

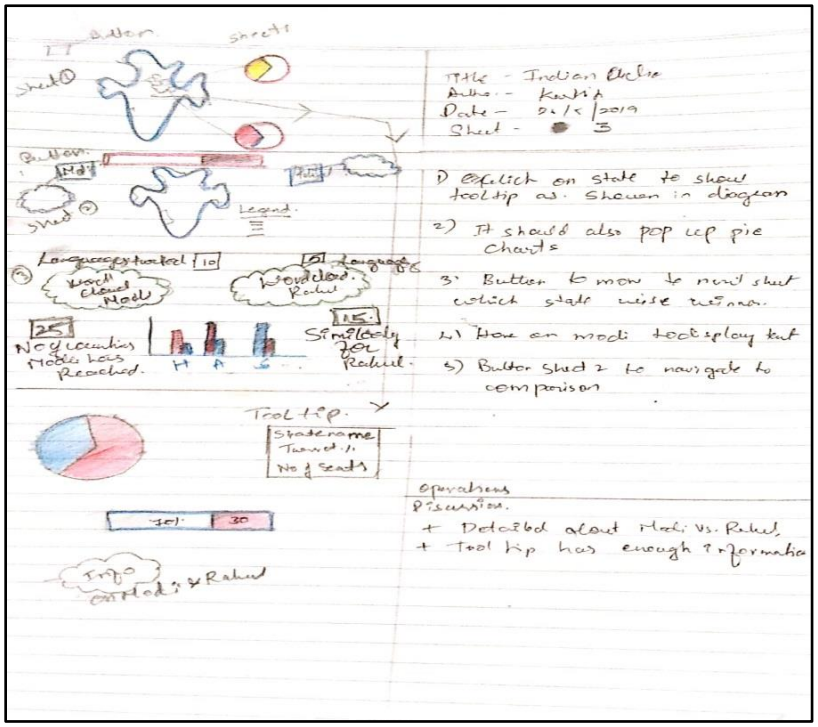
Design Sheet 1: Idea Generation



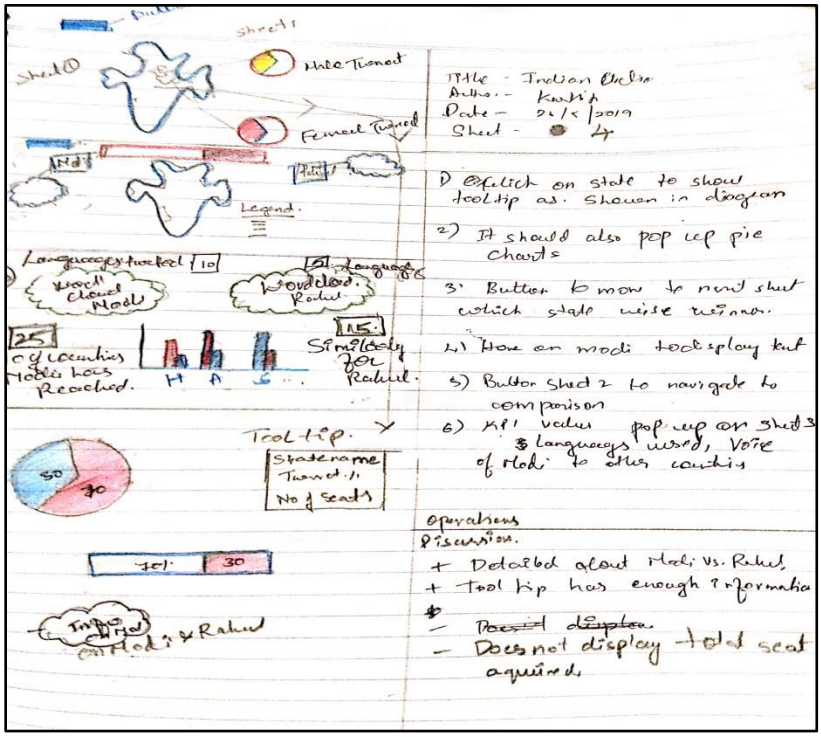
Design Sheet 2:



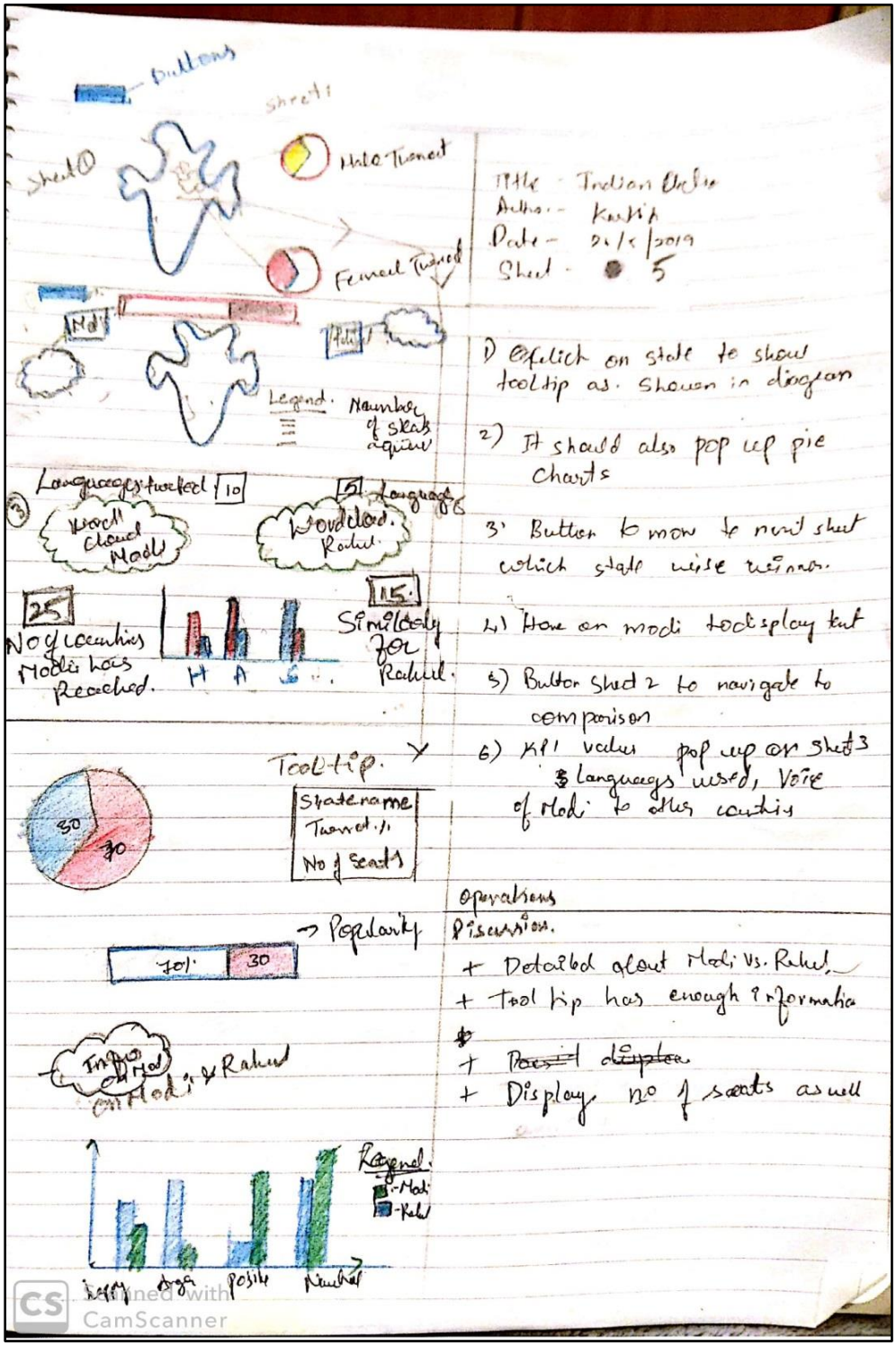
Design Sheet 3:



Design Sheet 4:



Design Sheet 5: Final Sheet (during Presentation)



Design Sheet 5- Updated