A View of India Elections Past and Present

Social and Historical - 2004 - 2014

MadhuJalan (71310073)

Raghuram Lanka (71310043)

Reva Maheshwari (71310071)

Executive Summary

Our project presents an analysis of elections in India. It presents two views, past and present.

Present View: Social Media is being used extensively in the current elections by both parties and candidates. India has 103 million social media users of which a large proportion are the youth. Social Media sentiment analysis is a great way to understand the sentiment amongst the youth.

The present view of the elections presents the social and media sentiment towards the major parties and candidates in the months of February and March prior to the start of the Elections. The highs and lows are labeled to show why this is the case.

Past View: Corruption, Gender and Voter Apathy are some of the key issues that India grapples with in the elections. The dashboard analyzes the data for the General Elections in 2004 and 2009 to show the difference amongst the various states on the number of court cases, number of female versus male candidates and the voter turnout by location.

We attempt to present the views and ideas amongst the indian population using social media. We did not try to predict results of the Indian Elections through our analysis.

Contents

Positive momentum for Mamata through the month of March

1 Mamata government introduces online file clearing system

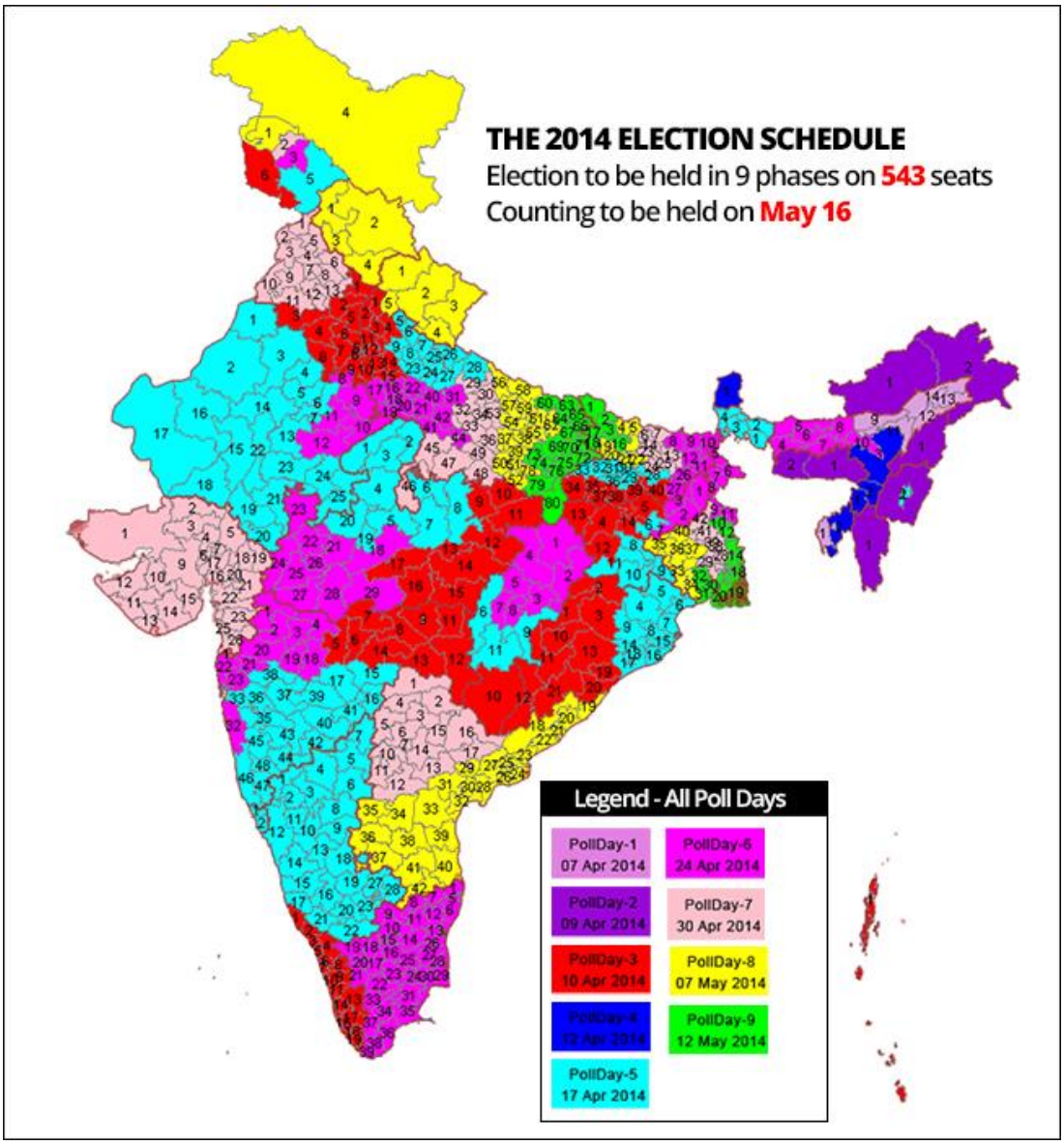
2 Anna backs off from Mamata rally in New Delhi

3 Mamata claims 'will emerge as 3rd largest party'

4 Division of West Bengal will not be allowed - Mamata

Background on Elections 2014

The world was watching the coming elections in India with great interest. The largest democracy of the world, with an electoral strength of 81.45 crores (814.5 million), went to the polls in 9 phases from 7th April, 2014 to 12th May, 2014. Voting took place in 543 parliamentary constituencies. The graphic shows the constituencies where elections were held in each phase.



Source: http://indiatoday.intoday.in

Though there are 6 major national parties, the two parties that were competing head to head are Bharatiya Janata Party (BJP) and Indian National Congress (Congress). Coalition politics were supposed to play a big part in determining who governs. Therefore, many of the regional parties

are also being watched closely. For the sake of simplicity we just show the sentiment for the 2 major parties and 2 major candidates.

The two leaders of the major parties that were competing to become the Prime Minisiter is Rahul Gandhi of Congress and Narendra of BJP.

Role of Social Media in Elections 2014

This is the first elections in India where social media gained extensive importance. It was not just the regular Indian tweeting, but also candidates, political parties and top leaders.

66 percent of India's population of 1.25 billion is below the age of 35 and almost 72 million Indians are in the age bracket of 18 to 23. There are 103 million social media users in India who are largely in the younger age bracket. A large chunk of these users use their mobile phones and smart phones to access the social media. The politicians have realized that this is the group they should be reaching out to through social media.

Sentiment Analysis

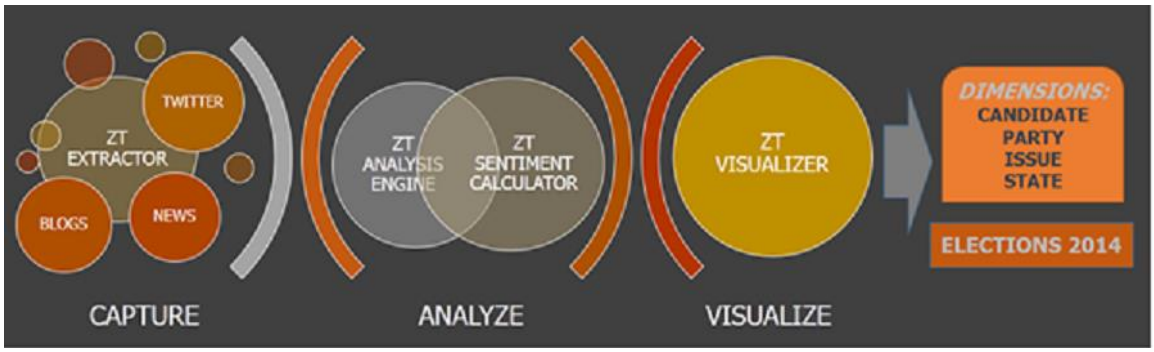
The rise of social media has provided us with a means of having conversations with family, friends, communities and the world at large. Forums such as Twitter are great resources for us to be able to follow conversations on any topic be it social, political, financial or business.

Sentiment analysis seeks to look at all the conversations and provide information on what is the general emotion amongst people on a particular topic, person and / or issue. While many levels of granularity can be provided, we look at positive, negative and neutral.

There are numerous factors to consider when providing an accurate representation of sentiment of any piece of data. Cultural factors, linguistic nuances and differing contexts make it extremely difficult to turn a string of written text into a simple positive or negative sentiment. Similarly one has to consider how to take the opinions of thousands and then aggregate them into a representative view.

We do not attempt to be exact because sentiment analysis algorithms at this point do not provide us with very accurate calculations, but provide general insights on the sentiment – positive, neutral or negative for a particular party or candidate. We also try to present some analysis on the issues may be related to the sentiment as well.

Technology



The analysis and various visualizations created can be found at http://www.ztanalytics.com.

We use a variety of tools to extract data. The sources of data Twitter, blogs and popular news media. Rather than attempting to extract all data we extract multiple samples. We apply different statistical techniques to analyze the data and categorize them into a Positive, Negative or Neutral sentiment for candidates and parties across different locations.

Most of our graphs show daily sentiment for particular parties and candidates. These have been correlated with news events to verify their accuracy and provide greater insights.

We used Python to write our code. Tableau and MS Excel were used to generate the visualizations. The site was hosted using MS Azure.

Data Sources

Date for 2004 and 2009 Elections (Past): The data is extracted from the open source shared by Election Commission of India on the website (www.eci.nic.in). The website contains plethora of data about elections. However, we would consider only the portion of data related to general elections across 2004 and 2009.

Some of the fields in the database are:

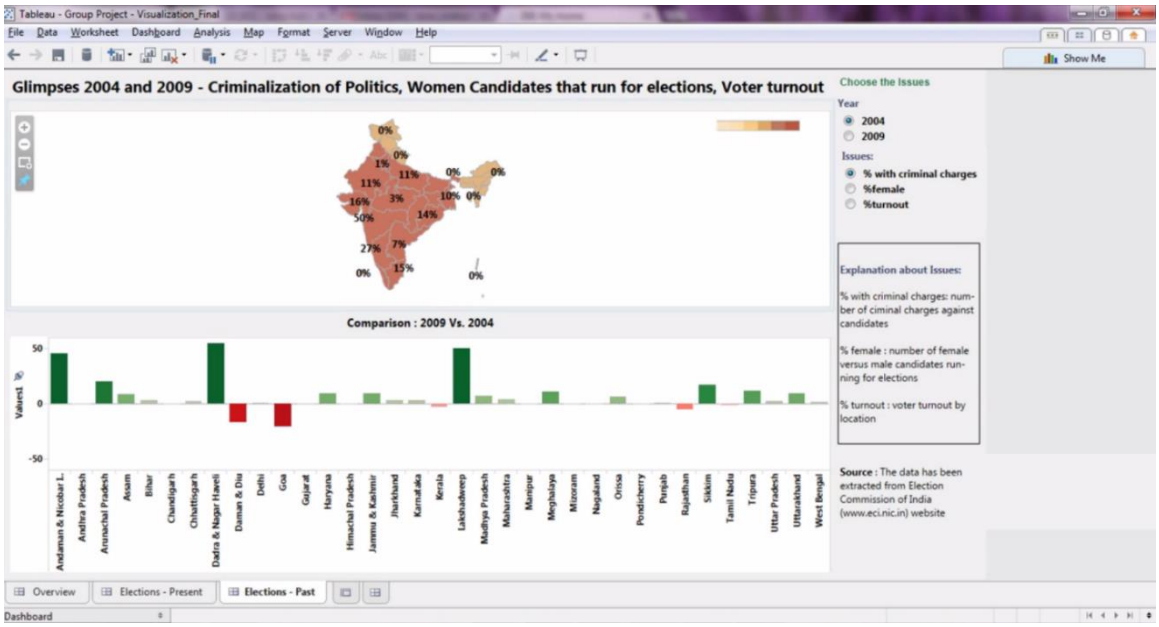
- Election Year
- Election Type
- Candidate ID
- Candidate Name
- Candidate Education
- Candidate Assets
- Candidate Gender
- State

- Constituency
- District
- Total Electors
- Total Votes
- Total Turnout
- Category
- Party
- Age
- Case Punished
- Case Type
- Total Cases Punished
- Total Cases Registered

Data for Elections 2014 (Present): For the current elections we will also look at popular sentiment for some of the current parties and candidates gathered through Twitter. Here we are using Python scripts to extract, transform, aggregate and conduct data analysis. A brief overview of the process is given below.

Past Elections

The objective here was to showcase trends on 3 major issues – Criminalization of Politics, Women in Politics and Voter Participation for the 2 past Indian Elections in 2004 and 2009 by State.



We were looking at 3 analytical aspects of the past elections:

1. The percentage of candidates by state that have criminal cases against them
2. Percentage of Women candidates by state that run for elections
3. Percentage Voter Turnout by state

The data was presented in two panels, the top panel shows a map where each of the above metrics is shown in a map of India. This data could be filtered for either 2004 or 2009. The bottom panel shows a deviation graph where we are seeing the percentage change in the numbers between the two years – 2009 and 2004.

This data is shown by location in a map, because given the huge regional differences in demographics, education, political preference in India, the differences between each of the states is an important analysis to undertake.

Data Transformation

There are over 12,000 records pertaining to 2004 and 2009 period. The data was cleaned, transformed and aggregated using MS Excel. Significant work had to be done to get it into the mode in which it would be visualized in easily digestible chunks.

Data Visualization

We chose to show this data by location in a map at the state level, because given the huge regional differences in demographics, education, political preference in India, the differences between each of the states is an important analysis to undertake. This data was visualized in Tableau.

The colors used here in the map are shades of brown. The reason for this is that the numbers are alarming and these colors are meant to draw attention to the fact that action needs to be taken in these areas. The bar graph shows green for positive improvement from 2004 to 2009 and red where the numbers have become worse between the past two elections.

Limitations:

- In order to preserve space, the labels are shown in a vertical instead of horizontal bar graph.
- To preserve simplicity while allowing for detailed analysis, we are not showing detailed information for each of the states.
- While we would have liked to show high level overview of each of these 3 aspects 'at a glance' in the dashboard, we found it was getting too cluttered.

Present Elections

Here the objective was sentiment for the 3 major parties and 3 major prime ministerial candidates for the ongoing India's General Elections 2 months prior to the start of the elections i.e. for February and March. These were Rahul Gandhi from the Congress, Narendra Modi from BJP and Arvind Kejriwal from AAP. We also did some analysis on some of the leading players such as Jayalalitha, Mamata Banerjee.

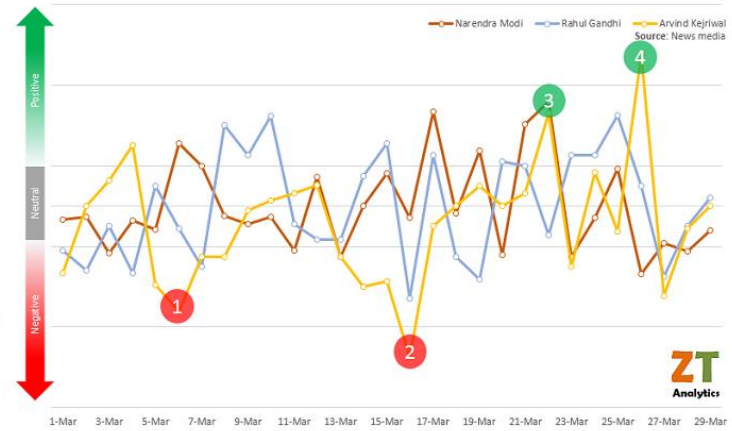
THe analysis was done along various dimensions for Media and Twitter. Some of them are presented below.

- Word Count of major players to show who was most talked about. This does not show positive or negative mentions.

NEWS MEDIA

TWITTER

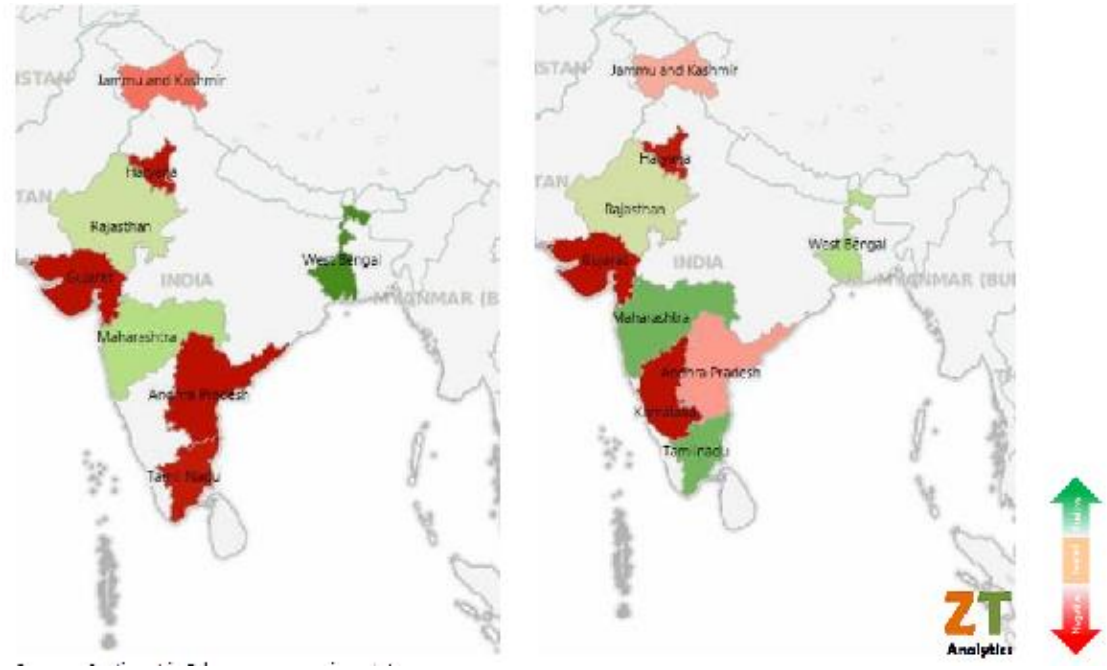
- Daily analysis of sentiment for various parties and candidates with the major points highlighted.



1. Mumbai Police lodge FIR against Kejriwal
2. Kejriwal insults media and then backtracks
3. News on INR surging to 35 – 40 against $ if Modi PM
4. Kejriwal roadshow brings Varanasi to a halt

- Sentiment of party and candidate across time

FEBRUARY



Congress

Rahul Gandhi

Data Transformation

Data transformation was a significant effort here. Some of the issues were:

- Candidates and parties were called by different names. In Twitter data often short forms (Narendra Modi and NaMo are the same person), vernacular and slang are present. All this needed to be filtered out and then given a common form.
- All sorts for emotion were characterized into positive, negative and neutral which is sometimes confusing.
- Rahul could mean Rahul Gandhi or Rahul Dravid so giving context to ensure we picked up the right tweets was important.
- Important to deal with duplicates.
- Understanding how to deal with retweets versus actual tweets.
- The importance of someone tweeting with a large following (Eg celebrities) versus those with a smaller number of followers.
- Twitter has a huge number of fake accounts. We had to figure out how to deal with what seemed like fakes.

Algorithms

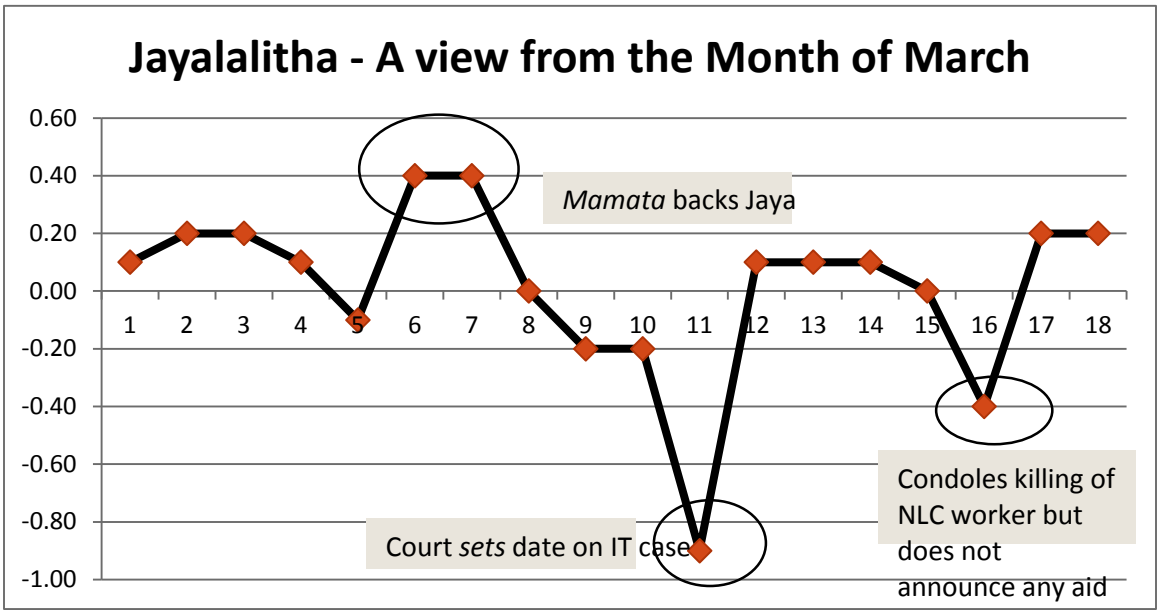
At the simplest level, we tried a count based algorithm – counting positive and negative mentions. A matrix of frequencies was calculated and then this was used to calculate sentiment. The words for sentiment were compared against a dictionary of positive, negative and neutral words. However this had several problems of inaccuracy.

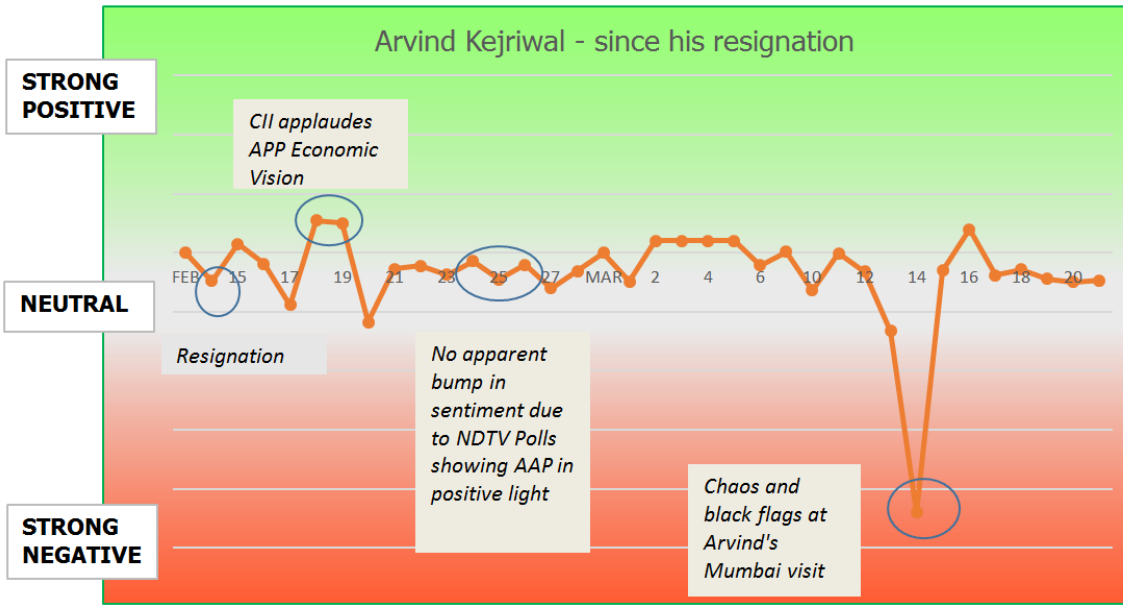
We then moved to using more sophisticated techniques. Some of them are discussed below.

- We eliminate objective sentences such as “I am going to Nevada.”, “I own a cell phone” etc. We only look at subjective sentences which convey some sentiment such as “I love my iPhone”, “Rahul Gandhi needs to grow up and be more rational” etc.
- We extract entities (iPhone, Samsung S3) and aspects (display, batter) and then associate sentiment with each of these aspects and overall with the entity.
- We used various algorithms such as Naïve Bayes and other Rule based classifiers to classify sentiment.
- We tried to make our algorithms self learning – for example we saw that Rahul Gandhi and Pappu were appearing together several times and Pappu was used in a derogatory sense.
- We looked for negations such as “This is not good”.
- We don’t just look for individual words but also n-grams.

Data Visualization

Data Visualization was a challenge here. We had a lot of data and a lot of insights, but to present it in a easily consumable manner which was aesthetically pleasing was quite tough. We experimented with various different kinds of graphs to figure out how to represent the data. Two of the rejected visualizations are shown below.

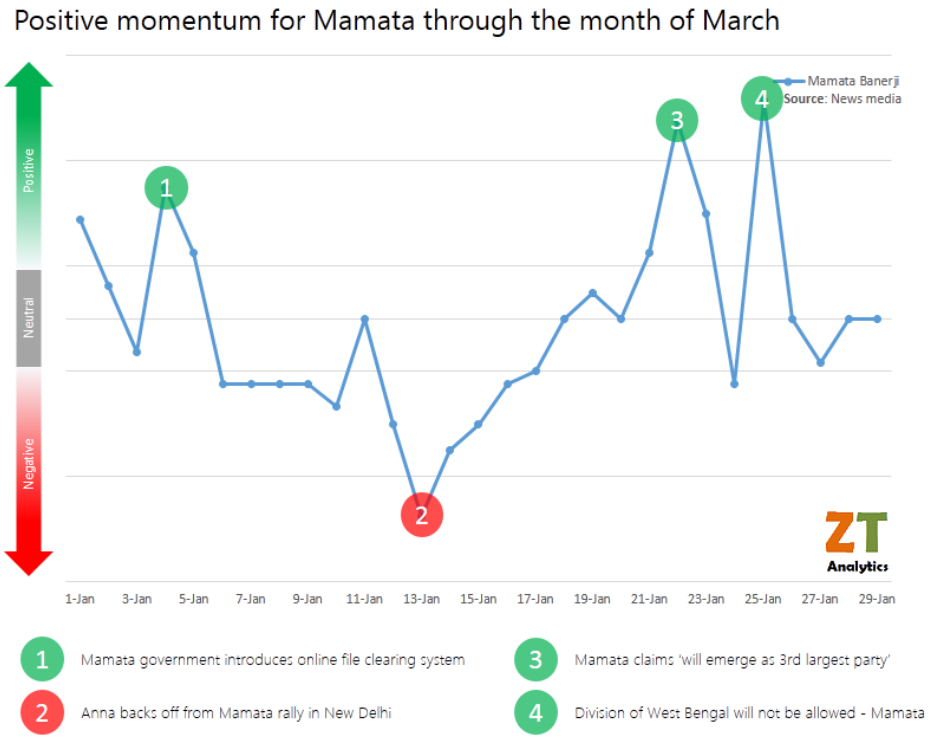




We chose line graphs to represent the sentiment because this was simple and the only one that provided easy comparison between two data points at the same point in time as well as be able to show some trend over a period of time for a particular candidate or party. Rather than trying to clutter the graphs with too many lines, we left plenty of white space so people could comprehend the information easily. We wanted to highlight the peaks and troughs as opposed to the general trend. We also showed the same information by location as the states in India

have very different personalities. We decided against putting a scale because we thought that would only confuse the user.

So finally chose the following format:



Conclusion

The aim of our analysis was to present data visually in a simple and intuitive way. We assume no technical understanding of sentiment analysis on part of the user to understand the data presented. Anyone interested in the elections and what opinions were held by people prior to the elections can use our website. The differences by states was interesting. Being able to compare media and social media reports was another interesting aspect of what we did.

Based on Facebook and emails we saw quite a few positive comments on the analysis we did. However, some people did not understand the idea of sentiment analysis. Others questioned the validity of using only media and social sources as India has a largely rural population that does not have access to social media.