Social Media Feed Analysis

Aditeya Baral
E L Mahima
Vinay Kirpalani

Introduction

Social Media

- Social Media has brought us closer than even before and has provided a common platform for us to communicate with one another.
- We live in a world where everyone is connected, in one way or another. Almost 4 billion people in the world are active on social media right now.
- Every second, more than 6000 tweets are tweeted, more than a 100 thousand pictures liked on Instagram and much more.

Why Social Media?

Although Social Media has its downsides, we absolutely cannot deny the potential it holds in various domains. From businesses to even politics, social media plays a pivotal role.

An average person will post about 45% of their life on social media, thus making it easier for us to perform an in-depth analysis of the demographic based on their activity.

Social media analytics is the practice of gathering data from social media websites and analyzing that data using social media analytics tools to make decisions. The most common use of social media analytics is to mine customer sentiment.

Twitter for Social Media Analytics

Twitter Feed Analysis

Twitter averages nearly 500 million tweets per day and more than 6000 per second!

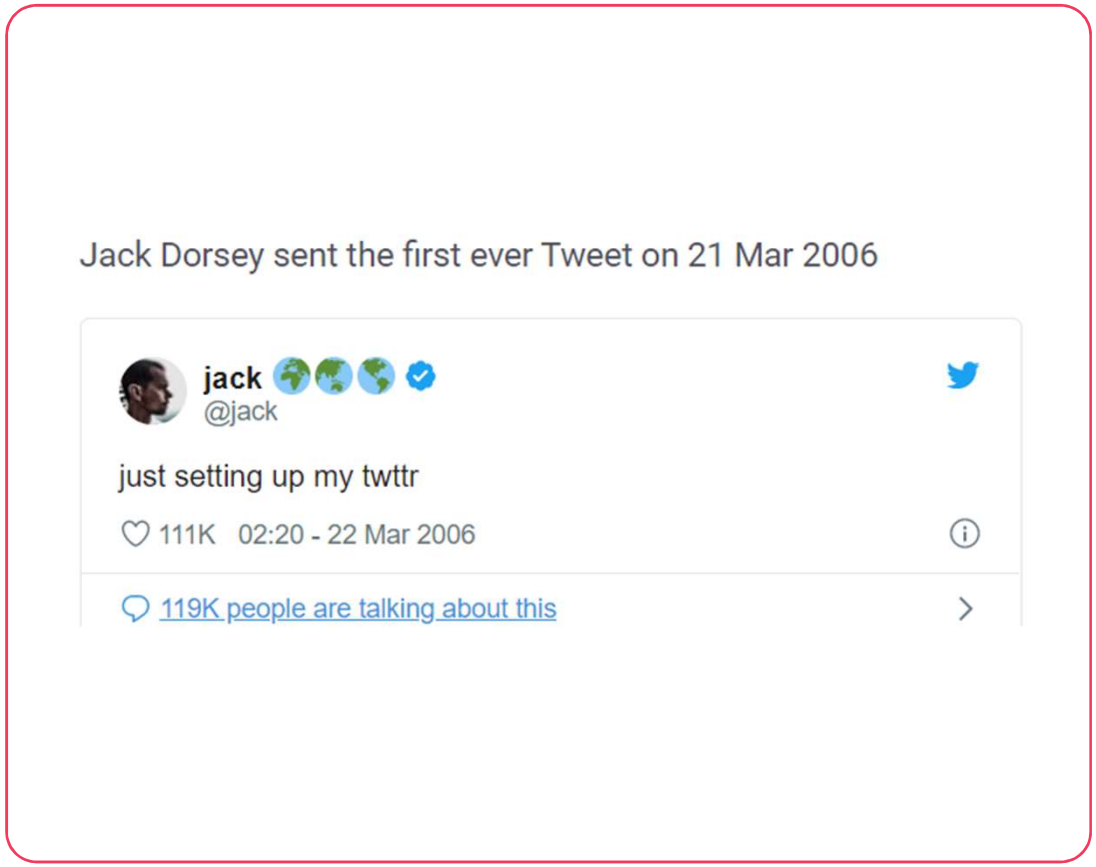
It has more than 330 million active users, and 1/10th of them are Indians.

Twitter provides a valuable tool for social media analysis because it is a never-ending resource which is self sufficient and keeps along with the times.

It provides a platform to post any kind of media and with a total character length of 280, it provides us with enough material to analyse sentiment accurately.

Multiple social media platforms have been created and many of them have been discontinued but Twitter still rides strongly with its large following.

Twitter Grew from this



To this!?


Bear Grylls
@BearGrylls
Tonight watch my journey with PM @narendramodi for Man Vs Wild on @DiscoveryIN - Together let's do all we can to protect the planet, promote peace & encourage a Never Give Up spirit. Enjoy the show!
#PMMODIONDISCOVERY
Discovery
MAN vs. WILD
WITH BEAR GRYLLS AND PRIME MINISTER MODI
TONIGHT 9PM
11:01 AM · Aug 12, 2019 · Twitter Web App


OBJECTIVES
STRATEGIC
VALUE
MISSION
STATEMENT
INTEGRITY
CONDUCT
SERVICE
INNOVATION
COMPANY
MARKETING
PRINCIPLES
PLAN
IDEOLOGY
INNOVATION
CUSTOMER
OBJECTIVES
VISION
RULE
GUIDING
PRICE
EXPENSE
COMMERCE
GOALS
EXCELLENCE

What are we doing?

Almost 400 million tweets were made during the Lok Sabha Elections held in 2019. Tweets were made in English, Hindi as well as other languages.

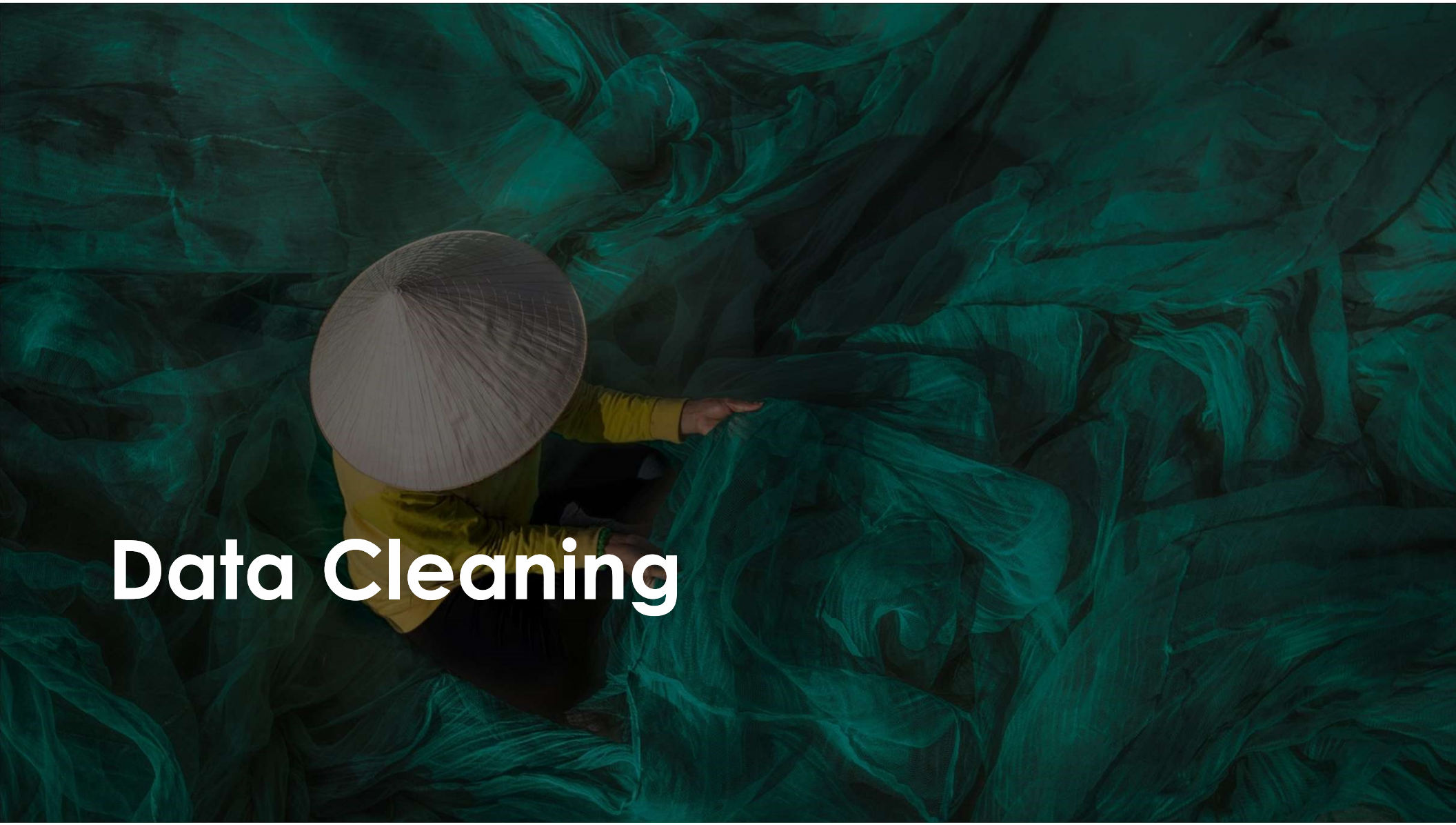
Our project aims to analyse all the 23rd hour tweets made between March and May and draw conclusions based on social media activity.

We have gauged people's thoughts and emotions on social media to predict the shift in support towards a party and compare the accuracy with actual poll results.

The Dataset

Features of the Election Dataset

- We tried to build our own dataset, but this was impossible as the Twitter API did not support tweet extraction for those which are more than 2 weeks old.
- The dataset used for this project is a subset of a much larger dataset from Kaggle. The reason for omitting a few columns is that they contained processed data based on other columns.
- The final dataset used by us had 12 attributes and nearly 46000 rows.
- The data had lots of noise and various inconsistencies which had to be cleaned.


Data Cleaning

How did we do it?

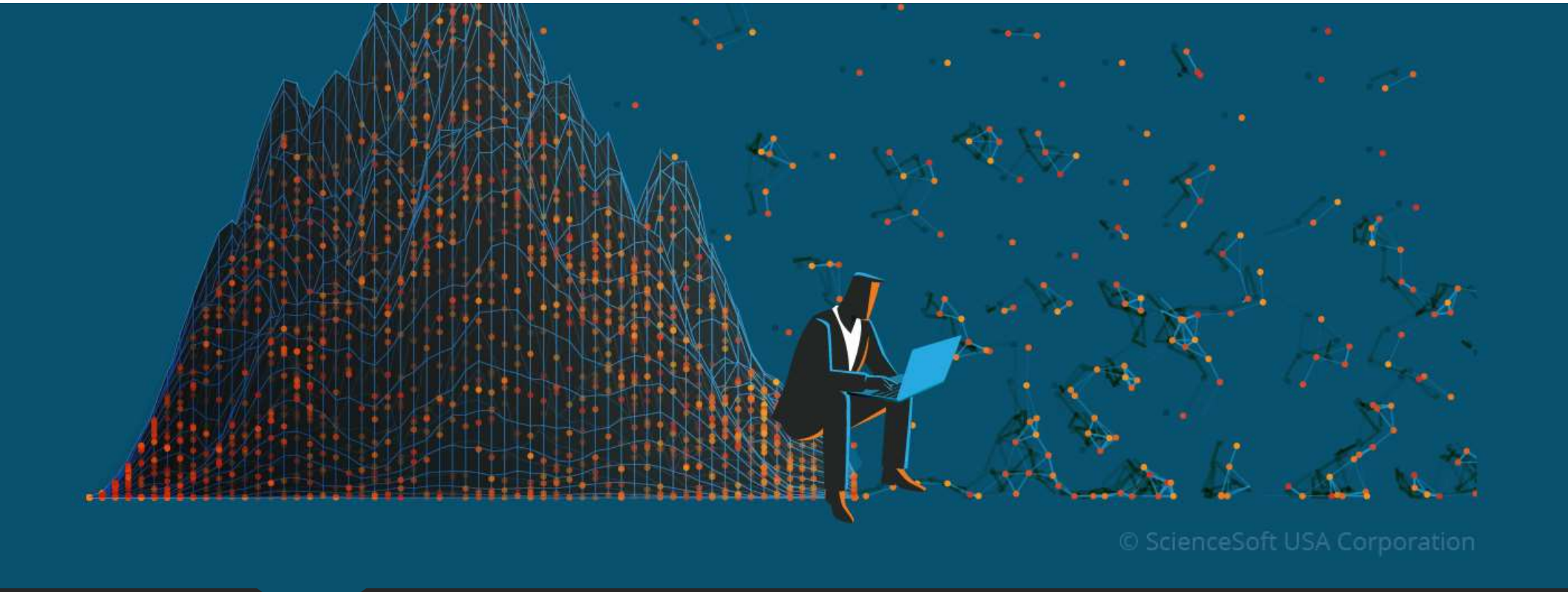
- Removed US Congress Tweets and tweets made after results.
- Replaced NaN values with default values.
- Converted times to datetime objects and sorted the dataset.
- Mapped missing cities and states and fixed name inconsistencies.
- The final dataset had 39 thousand rows compared to the original 46 thousand.

Comparison – Original Dataset

	D	E	F	G	H	I	J	K	L	M	N	O
1	**user_location**	**full_text**	**quote_count**	**reply_count**	**retweet_count**	**favorite_count**	**hashtags**	**user_mentio**	**City**	**State**	**District**	**Country**
2	Guwahati	lot of prominence in	0	4	113	113	ActEast	narendramo	Guwahati	Assam		India
3		RSS in school days itself?	0	0	5	5						India
4	New Delhi	Vaccum!	0	4	31	31			New Delhi Municipal Council	Delhi		India
5		India's #Icecream Industr	0	0	3	3	Icecream					India
6	Delhi | Odisha	in Pulwama Attack	0	0	9	9				Delhi		India
7	New Delhi	@ncbn gave minority to	0	0	18	18		ncbn	New Delhi Municipal Council	Delhi		India
8	Others	awareness on PM Sh.	0	0	12	12		narendramodi				
9		A multitude of supporters	1	6	317	317						India
10	New Delhi	court latest by10.30			0	0			New Delhi Municipal Council	Delhi		India
11	https://wwwface	name for herself as one	1	0	3	3						
12		#BCCIBetraysBraves | W	0	10	34	34	BCCIBetrays	Braves			India
13	Maharashtra	let's look back and see	0	0	0	0				Maharashtra		India
14	Odisha	Odisha is blessed with so	1	10	120	120			Vidisha	Madhya Pradesh		India
15	Banglore	@mepratap and	1	6	61	61	EveryVoteFo	mepratap,na	Bangalore	Karnataka		India
16	Bangalore	and was waiting to	0	0	1	1			Bangalore	Karnataka		India
17	Others	@Mayor17Ian @marklev	0	1	0	0		Mayor17Ian,	marklevinshow			
18	à¤ªà¤‚à¤œà¤¾à¤	through the night to seal	0	2	47	47						
19	Karnataka	Over hyped person!!!	0	0	1	1		ragusmg,pra	Kanapaka	Andhra Pradesh		India
20	Others	This is the article by Luke	7	11	141	141						
21	Washington DC	organizations take	12	18	168	168						
22	Others	Congress & the entir	0	0	1	1						
23	Others	Modiji's Balakot airstrike	0	1	6	6						
24	San Francisco	million over the last	28	35	1426	1426						
25	Others	Samjhauta & Malega	0	7	276	276						
26	New Delhi	@RahulGandhi joins	21	181	3720	3720		RahulGandhi	New Delhi Municipal Council	Delhi		India

Final Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	last_updat	tweet_id	created_a	full_text	processed	Party	quote_cou	reply_cour	retweet_c	Importanc	hashtags	Neutral	Positive	Negative	Compounc	user_ment	City	State	Country
2	########	1.09E+18	########	@DasCom	dascomrac	BJP	0	0	1	0		0.625	0.375	0	0.4588	DasComra	Guwahati	Assam	India
3	########	1.09E+18	########	Congress r	congress n	BJP	47	72	2353	1		0.868	0	0.132	-0.5719		Mumbai	Maharasht	India
4	########	1.09E+18	########	BJP has be	bjp win ba	BJP	1	12	513	1		0.582	0.276	0.142	0.5719		Kolkata	West Beng	India
5	########	1.09E+18	########	@inclusiv	inclusivem	Other	0	0	2	0		0.884	0.116	0	0.4404	inclusivem	Aligarh	Uttar Prad	India
6	########	1.09E+18	########	Senior Cor	senior con	Congress	0	1	159	1		1	0	0	0	ManishTev	New Delhi	Delhi	India
7	########	1.09E+18	########	Is it b/c th	bc congres	Congress	0	2	34	0		1	0	0	0				India
8	########	1.09E+18	########	#RahulLie	rahulliecau	Congress	22	34	680	1	RahulLieCa	0.879	0	0.121	-0.296		Mumbai	Maharasht	India
9	########	1.09E+18	########	Expect a m	expect me	Congress	1	11	203	1		1	0	0	0				India
10	########	1.09E+18	########	These	stori realli	Other	1	0	10	0		0.926	0	0.074	-0.25				India
11	########	1.09E+18	########	If there is	issu proble	BJP	3	3	123	1		0.847	0	0.153	-0.4019		New Delhi	Delhi	India
12	########	1.09E+18	########	#ForTheFir	forthefirst	BJP	0	0	87	1	ForTheFirs	0.901	0.099	0	0.3182		Gandhinag	Gujarat	India
13	########	1.09E+18	########	#ModiUn	modiunsto	BJP	2	3	109	1	ModiUnsto	0.803	0.085	0.111	0		Pune	Maharasht	India
14	########	1.09E+18	########	The pity	piti state a	BJP	0	0	4	0		0.913	0.087	0	0.2023				India
15	########	1.09E+18	########	Jim Jordan	jim jordan	Congress	2	1	61	1	WhitakerH	0.544	0.286	0.17	0.4019				India
16	########	1.09E+18	########	Rahul	rahul rahul	Congress	19	33	717	1		1	0	0	0	manoharp	New Delhi	Delhi	India
17	########	1.09E+18	########	Modi Govt	modi govt	BJP	0	2	48	1	PMUY	0.71	0.198	0.092	0.3818		Una	Gujarat	India
18	########	1.09E+18	########	BJP MP Sh	bjp mp sh i	BJP	58	225	7894	1	AbkiBaar4	1	0	0	0	ianuragtha	New Delhi	Delhi	India
19	########	1.09E+18	########	Prime Mini	prime mini	Other	52	295	10761	1		0.89	0	0.11	-0.4767	narendramodi			India
20	########	1.09E+18	########	#RafaleSca	rafalescan	BJP	1	1	37	0	RafaleScar	0.72	0	0.28	-0.5423	SaugataRo	Kolkata	West Beng	India
21	########	1.09E+18	########	Howard U	howard un	Congress	16	15	121	1		1	0	0	0				India
22	########	1.09E+18	########	#ForTheFi	forthefirst	BJP	0	0	8	0	ForTheFirs	0.696	0.145	0.159	-0.128		Andada	Gujarat	India
23	########	1.09E+18	########	How India	india defer	BJP	52	318	820	1	DefenceM	1	0	0	0	PMOIndia	New Delhi	Delhi	India
24	########	1.09E+18	########	@INCIndi	incindia ra	Congress	0	0	15	0		0.8	0.2	0	0.3612	INCIndia,R	Mumbai	Maharasht	India
25	########	1.09E+18	########	@ErikaCal	erikacalvo	Congress	0	1	4	0		0.59	0.311	0.099	0.7003	ErikaCalvo,mitchellvii,	AOC		India
26	########	1.09E+18	########	The Shutd	shutdown	Congress	0	0	1	0		0.909	0.091	0	0.2023				India



Working with Data

Processing Dataset

We needed to clean the tweets before analyzing them.

The tweets were preprocessed

Each tweet was stripped of any kind of numbers, hashtags and emojis/symbols and all contracted words were expanded using RegEx.

The tweets were finally stemmed using the Snowball Stemmer (Porter 2).

Analysing Sentiment and Importance

The polarity of tweets was measure using nltk's sentiment analyser. Its positive, neutral, negative and compound scores were tabulated and stored.

The number of replies, retweets and quotes defined the importance of a tweet. After using the descriptive statistics to decide the cutoffs, the importance was stored as binary coded 2 states.

Scoring Tweets

Scoring Tweets

We assigned an overall score to every tweet.

This score accounted factors such as retweets, quotes, replies and the previously calculated importance.

We defined a simple linear homogenous function in terms of factors with coefficients equal to the correlation between them and the positive sentiment.

The final score hence gave a weighed popularity measure of the tweet which depended on the factors.

Classification into Parties

Classifying Tweets into Parties

1

Since there was no tagged dataset available, we couldn't use any classification algorithm.

2

We hence went ahead with a simple frequency-based classification technique.

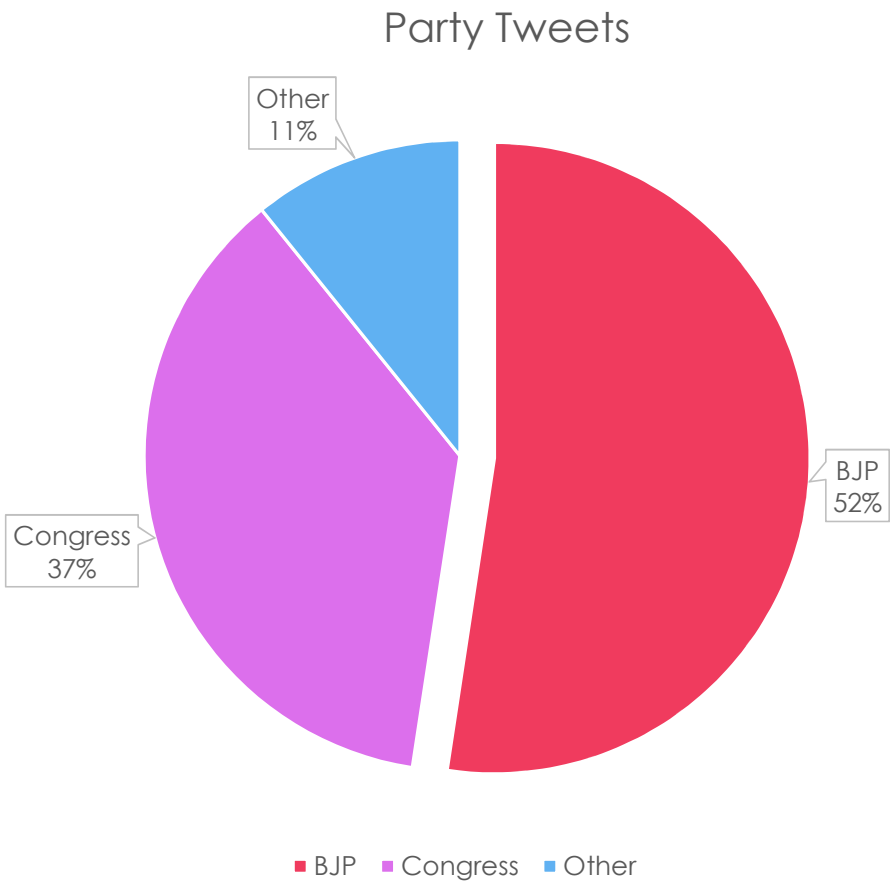
3

We distinguished between tweets by using the frequency of a set of terms related to different parties.

4

This method although slow, worked to a high degree of accuracy when manually checked

Party Distribution



Popularity

Popularity Scores

We obtained a regular time interval for days, weeks, fortnights and months.

We plotted the cumulative popularity for each party for against the time intervals. We found that the daily popularity score was the perfect scale.

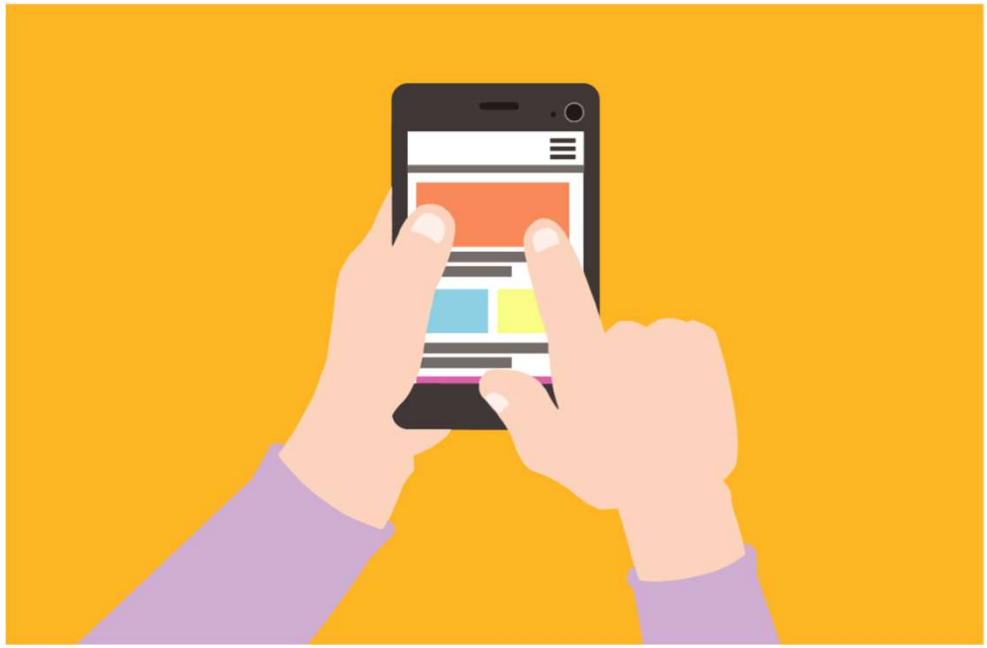
We plotted the popularity of each party in the top 10 states and the final popularity index of each party.

The plots obtained are an accurate representation of the actual shift in popularities of each party throughout the interval.

Graphs and Visualizations

Socially Active Regions

- We decided to find out which states tweeted the most during the General Elections.
- This helped us analyse whether all states had an equal say in a party's popularity.
- We obtained the states that had atleast 150 tweets and plotted a bar graph.
- However we were dissatisfied with the plot and felt that it could be made more interesting.



Geo Spatial Plots

Why Geo Spatial?

Since the data plotted consisted of regions, we felt that it could be better represented on spatial plot.

We decided to use a Geo-Spatial Choropleth plot for showing scores of different regions.

We obtained the India shapefile and converted each field in the dataframe to a geo dataframe using geopandas.

The states were then plotted by converting them into polygonal objects from the shapefile.

A value in the choropleth defines the depth of colour used, so darker shades represent higher values.

Hashtags

#WhatDidWeLookFor?

1. We decided to find the most trending hashtags during the General Elections
2. These metrics add on to our analysis and help us find out the support a party receives on social media
3. #BJP, #Modi and #LokSabhaElections2019 were some of the most trending hashtags
4. BJP related hashtags were found to be more than Congress related ones
5. We also analysed the distribution of hashtags across different states
6. BJP hashtags came from all over the country while Congress hashtags came only from Congress dominant states


Tagged Handles

@what_did_we_do?

We also analysed the number of times people decided to tag the leaders themselves in their tweets

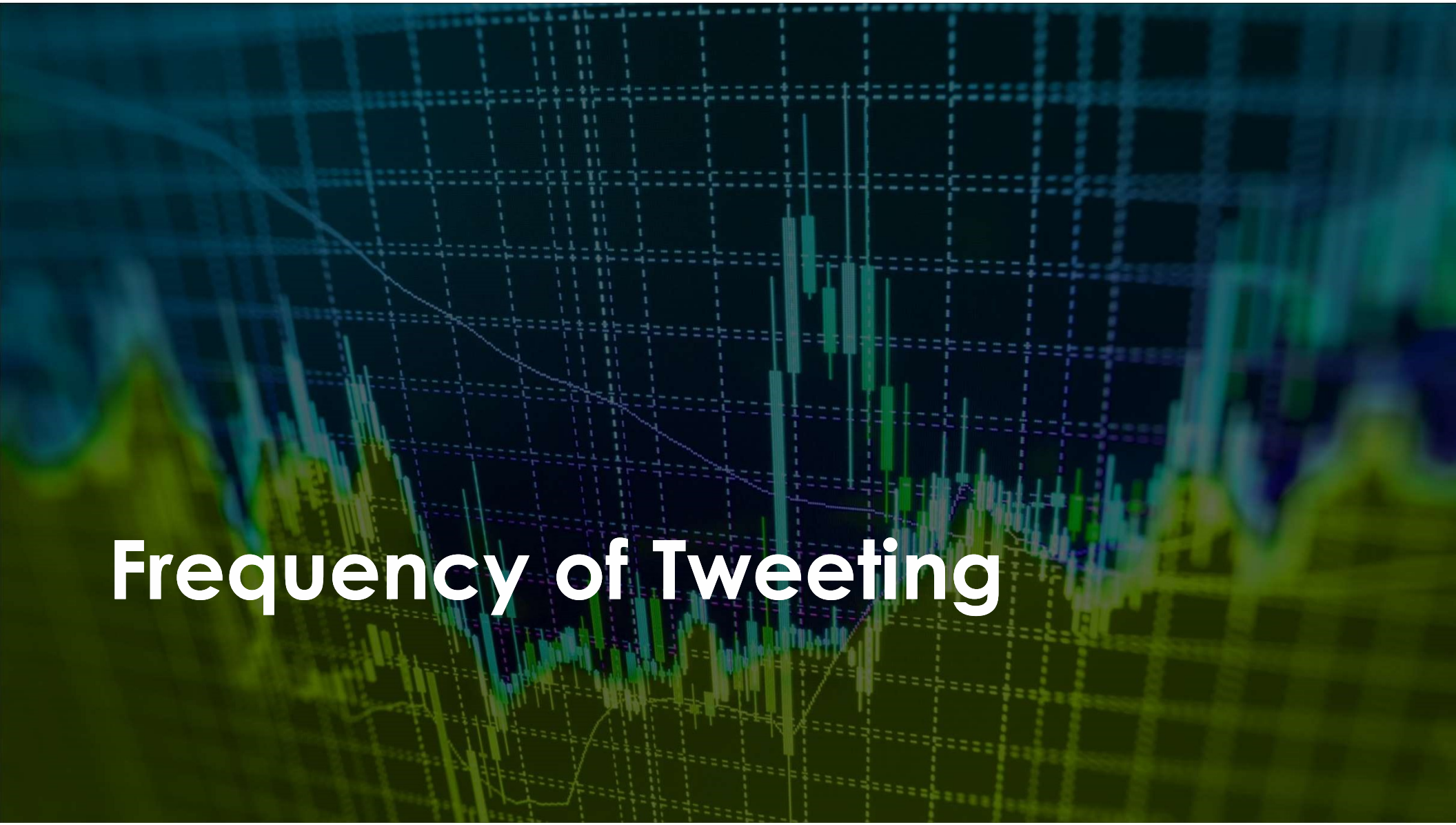
We found the distribution to be non uniform; only 4 handles were tagged a reasonable number of times

BJP, Narendra Modi, Congress and Rahul Gandhi were tagged in different tweets

Ironically, people tagged Congress and Rahul Gandhi twice as frequently as BJP

However, Modi seemed to be an extremely popular choice for tweeters as he was tagged more than the other three handles combined

Towards the end of the timeline, the frequency of tagging handles decreased drastically


Frequency of Tweeting

How often did they tweet?

1. The frequency of tweets did not remain the same over the time period
2. The frequency of tweeting was more than 800 tweets halfway through the timeframe
3. There were two dips in in the timeframe – right after beginning, after the first few days of voting and towards the end
4. There was a drastic fall in the number of tweets and even went down to almost 300 just before the declaration of results
5. The average number of tweets during the middle were about 600 tweets
6. The number of tweets dropped to 500 on 22nd May

Chances of Winning

Analysing Chances of Winning

- We analysed each party's chances of winning leading up to the results
- To do this we plotted the probabilities of each party's winning chances
- Coalitions between parties were also taken into consideration and probabilities of winning were plotted
- The probability scores were accurate and showed that BJP had a much higher probability of winning compared to other parties
- The plot also showed that parties except BJP and Congress barely stood a chance in the Elections

Poisson Distribution

The mean of the scores of each party were found per day

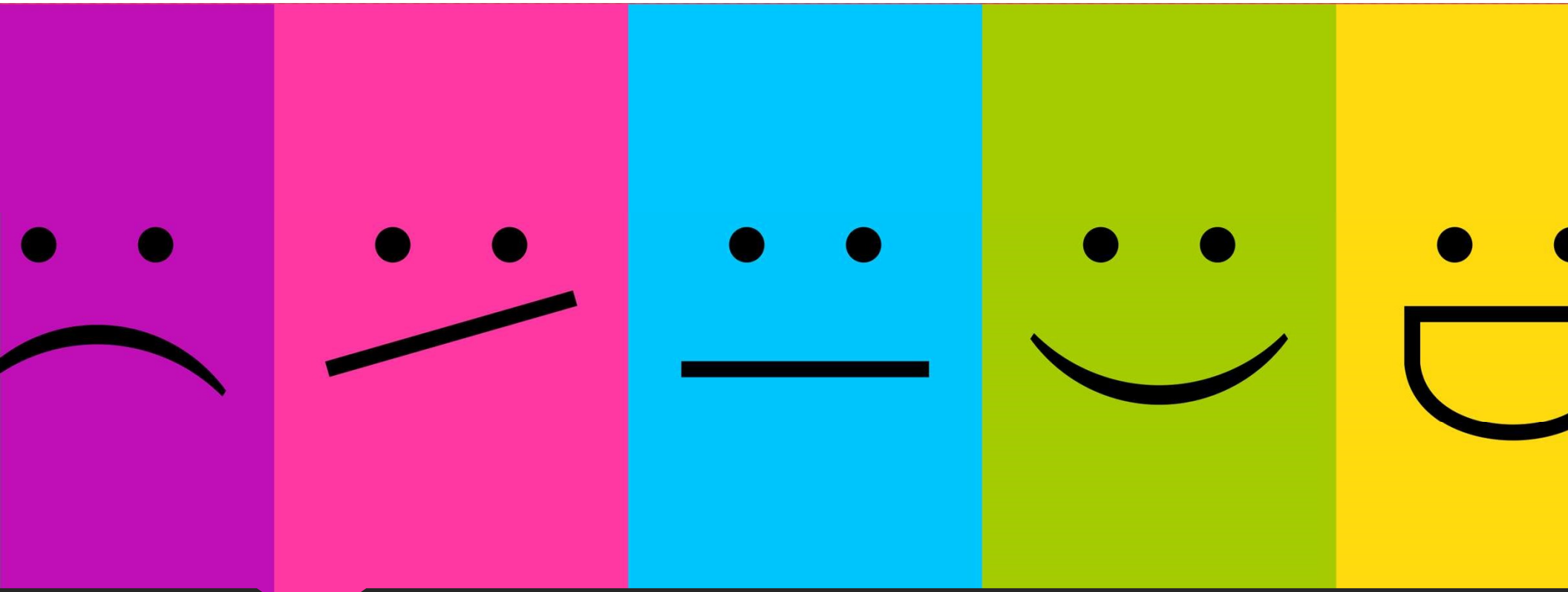
These were again averaged per day, and the corresponding λ calculated

The Poisson Distribution for each party was plotted

The curve resembled the distribution for λ = 1 for all parties

Since the means differ by microscopic values, they aren't reflected on the plot

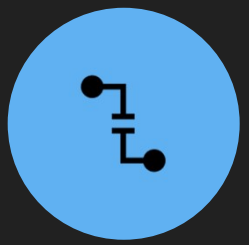
However, results of the distribution show that BJP is the clear winner



Emotion Plots

Measuring Moods

We also analysed the emotions and moods associated with every party and state



The tagged NRC Emotion Lexicon dataset was used to do this

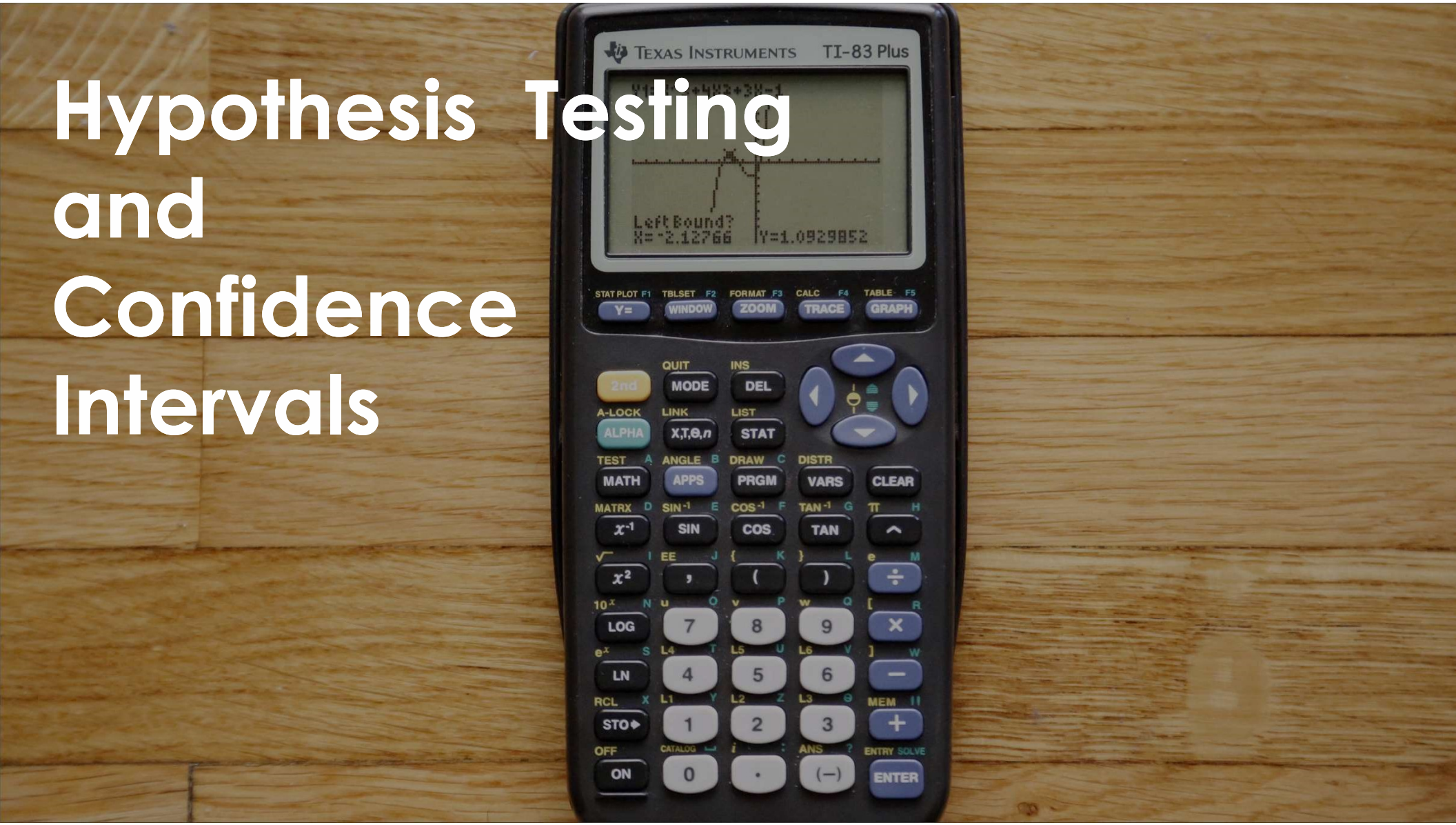
This helped us find out what the people felt about the parties and the moods associated with them

The metric provided a certain degree of measure beyond just the positivity and the negativity of the tweet and helped us find out more about what people think

What did people feel?

- BJP had more fear, anger and trust associated with it, along with its popularity
- The distribution in moods across states was uniform, suggesting that most people felt the same way irrespective of where they stayed
- Feelings of disgust and surprise measured lower values compared to anticipation, trust, joy and fear
- The positive and negative emotions also matches the popularity scores of each party

Hypothesis Testing and Confidence Intervals



Hypothesis Testing

We performed two hypothesis tests – one on Score and the other on Popularity

The first was an upper tail test to see if the mean score is greater than 25

The latter was a two-tail test to see if the Popularity is not equal to 0.05

In both tests, we used a 95% confidence interval and the null hypothesis was rejected

Normal Probability Plot

Normal Probability Plot

- We assigned ranks to all the values from i = 1 to n
- Their cumulative probabilities were then calculated as $\frac{i-0.5}{n}$
- Their Z Scores were calculated and plotted vs their original values
- A straight line plot was obtained which showed that the values were normal

Chi Squared Test Statistic

Testing for Chi Square

1

We use the Chi Square Test to check if two columns are independent

2

It was applied on the sentiment polarity scores – positive, neutral and negative

3

We found the Chi Square to be less than 10%, which hence showed that they are not independent of each other

4

From the Chi Square value we can say that the region isn't plausible

THANK YOU!