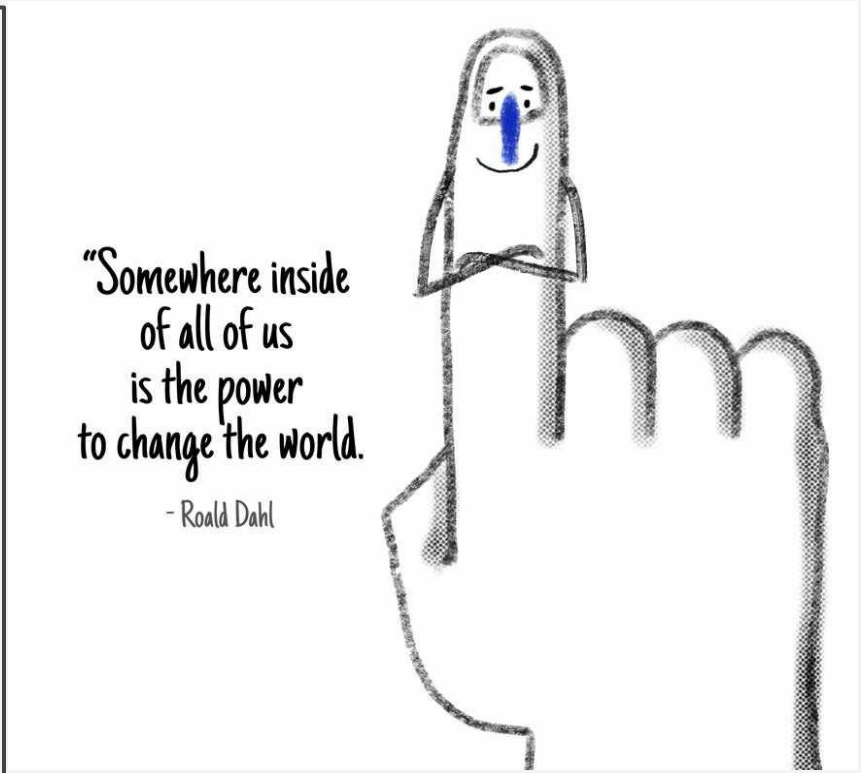
SOCIAL MEDIA ANALYTICS FOR

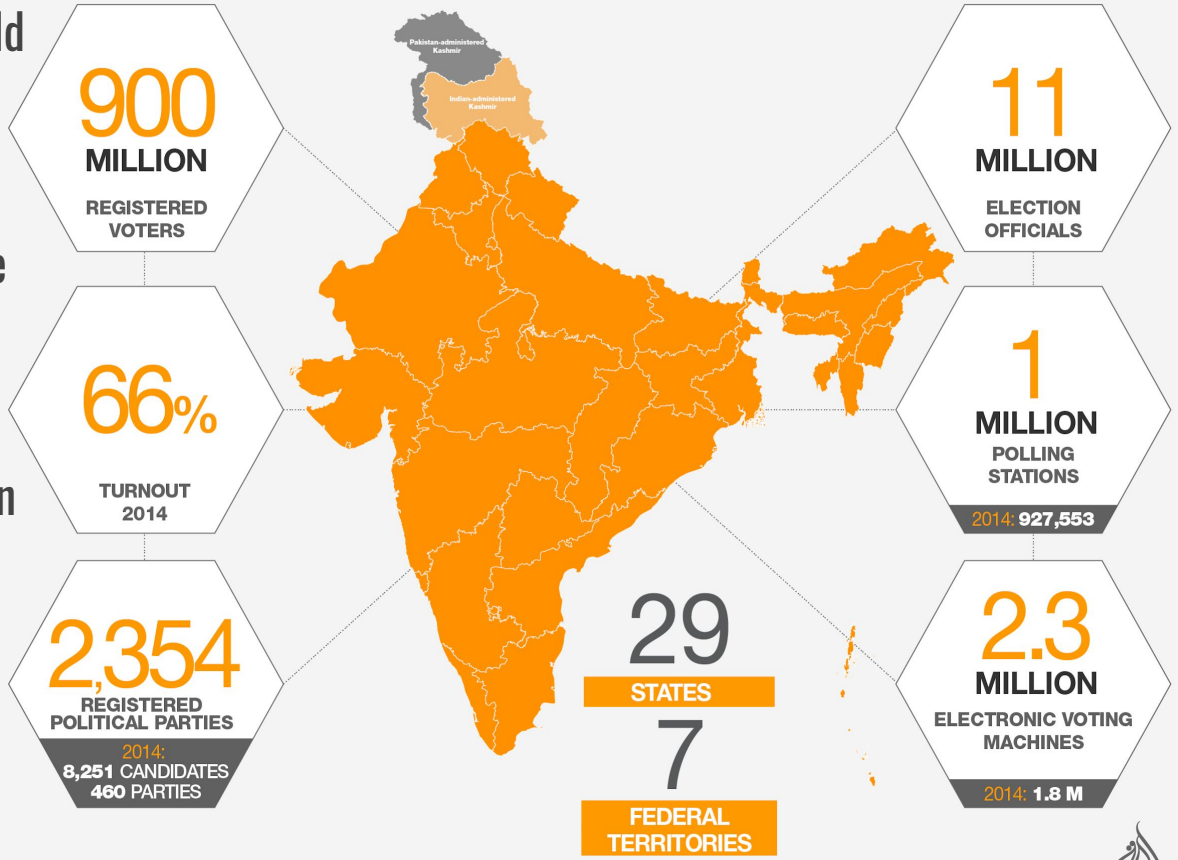
PREDICTING ELECTIONS

OVERVIEW



Elections

- Largest democracy in the world
- More than 900 million eligible voters
- 2019 had highest participation by women voters



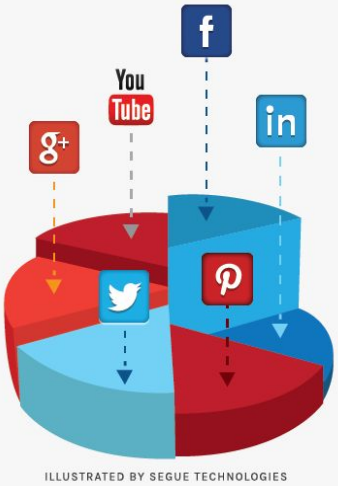
INTRODUCTION

- Importance of social media in modern age
- Twitter has approximately 100 million active users on daily basis.

- Twitter and Facebook for political communication, interacting with voters and promotions.
- Increased use of social media by political candidates has been reflected by the 2011 general elections in New Zealand.

Social media Analytics

- Social media analytics is the process of gathering and analyzing data from social networks such as Facebook, Instagram, and Twitter.
- Data offered by social media sites that gives insight into what people are responding to and engaging with on your social channels.
- Twitter - Twitter Analytics
  Facebook - Insights tab of Facebook pages
  Instagram - Facebook Insights platform



Sentiment Analysis

- Sentiment analysis is the process of determining the emotion underlying a bunch of words.
- It helps to understand the attitudes, opinions and emotions expressed within an online mention.
- Predominantly used in data science for analysis of customer feedback on products and reviews.

- Input:

  A document d

  A fixed set of classes C = {c1,c2,..,cn}

- Output: A predicted class c ∈ C

  Training set of n labeled documents looks like: (d1,c1), (d2,c2),...,(dn,cn) and the ultimate output is a learned classifier

Remove Stopwords("I", "Are", "Am", etc.)

Document means tweets, phrases, parts of news articles, whole news articles, a full article, a product manual, a story, etc.

"I ***love*** *this movie! It's* ***sweet****, but with* ***satirical*** *humor. The dialogs are* ***great*** *and the adventure scenes are* ***fun****. It manages to be* ***romantic*** *and* ***whimsical*** *while laughing at the conventions of the fairy tale genre. I would* ***recommend*** *it to just about anyone. I have seen it several times and I'm always* ***happy*** *to see it* ***again*****....."**

great	2
love	2
recommend	1
laugh	1
happy	1
.	.
.	.
.	.

document	w1	w2	w3	w4	...	wn	sentiment
d1	2	1	3	1		1	positive
d2	1	5	5	5		1	negative
d3	3	8	6	8		2	positive
d4	2	5	1	5		3	positive
d5	3	0	3	0		0	negative
d6	0	0	0	0		0	negative
d7	2	0	0	0		0	positive
d8	9	2	9	2		2	negative
...	...	...	...	...	...	...	...

Corpus

Social Media for Elections

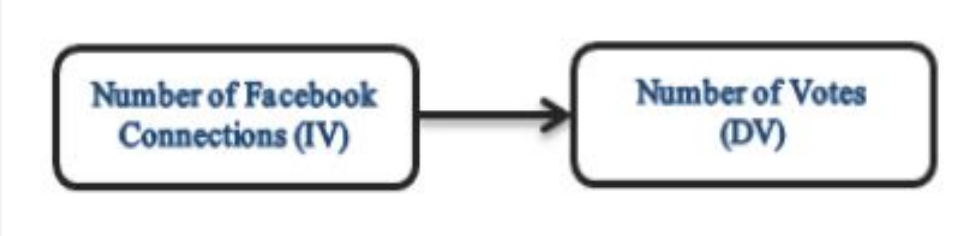
" Political parties and their IT cells are leveraging big data analytics to reach out to the voters with appropriate messaging in General Elections 2019 "

Creating Trending Narratives

Deciding Candidature

Methods

Method 1



- Number of votes was taken as a dependent variable and number of Facebook connections, along with other factors were taken as independent variables.

- The results reported the size of the network and the chances of a win are significantly correlated to each other.

Method 2

Volumetric and Sentiment Analysis

- Utilized by authors to evaluate the predictive power of Twitter data for inferring electoral results for three countries, Pakistan, India, and Malaysia.
- The data preprocessing was performed on approximately 3.4 million Tweets collected using Twitter streaming API.
- 90% of the Tweets from Pakistan and India are English, but on the other hand only 23% of the tweets were in English.

Volumetric Analysis

$$Vol_x = \frac{c_x}{\sum_{j=1}^{n} c_j}\% \qquad (1)$$

- Vol_x for any party x represents the volume of tweets and c_x represents tweets count.

Sentiment Analysis

$$Sent_t = \{\, 1, pos_t > |neg_t| - 1, pos_t < |neg_t|, pos_t = |neg_t| \,\} \qquad (2)$$

- $Sent_t$ is for sentiment of tweets pos_t represents positive tweets where as neg_t represents negative tweets.

- The results reported that the Twitter data was not effective for making election predictions for Malaysia, but in the case of Pakistan and India, it appeared as an effective and efficient for electoral predictions.
- By combining multiple techniques the proposed model for predicting electoral outcomes was also effective for candidates and parties having small vote count

Method 3

Senate Vote = *f* (partisan voting index + incumbency + participation advantage)

USING FACEBOOK

- Senate vote is the percentage of forecasted votes won by either two of the major parties
- Partisan Vote Index (PVI) is the past election results
- The metric from Facebook was used to calculate the incumbency and participation advantage. Incumbency is the holding of an office or the period during which one is held.
- The components attached with the participation variable include likes, active post engagements, and time slices. The track of fans and the post engagements has been continuously kept by the Facebook pages.

Method 4

Using machine learning - Naïve Bayesian approach

- Been proposed to predict the electoral result for the US presidential elections 2016
- Twitter tweets have been collected over the period of Three months, from December till February
- After doing simple preprocessing of data the model for sentiment prediction has been reported to achieve 95.8% accuracy.

USING TWITTER

Conditional Probability

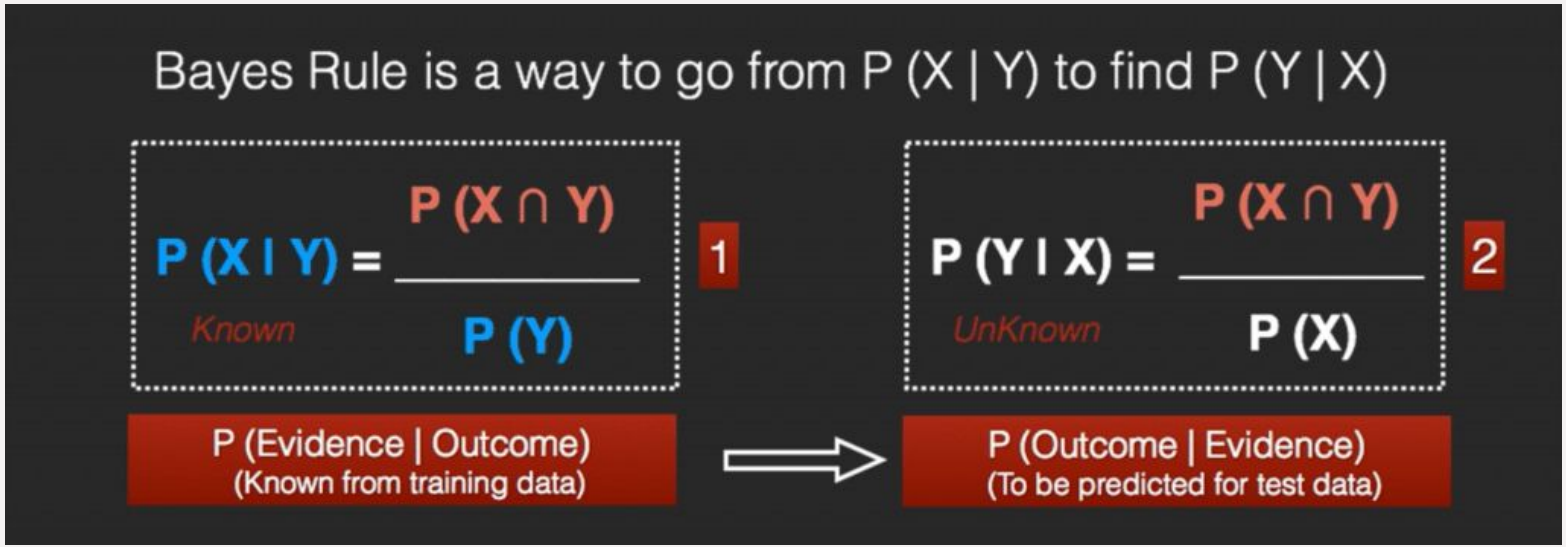
	Female	Male	Total
Teacher	8	12	20
Student	32	48	80
Total	40	60	100

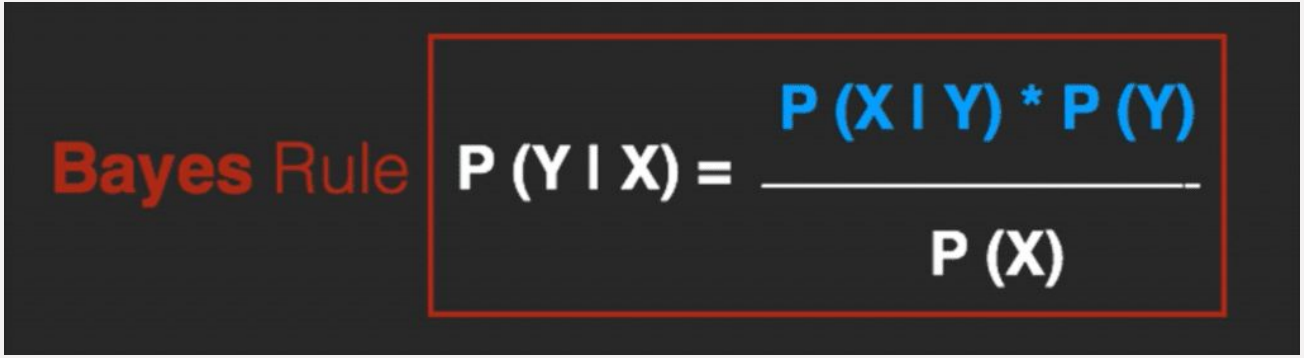
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

$$P(Teacher \mid Male) = \frac{P(Teacher \cap Male)}{P(Male)} = 12/60 = 0.2$$

Bayes Rule


Bayes Rule is a way to go from P (X | Y) to find P (Y | X)
P (X I Y) = P (X ∩ Y) / P (Y)
Known
1
P (Y I X) = P (X ∩ Y) / P (X)
UnKnown
2
P (Evidence | Outcome)
(Known from training data)
P (Outcome | Evidence)
(To be predicted for test data)


Bayes Rule
P (Y I X) = P (X I Y) * P (Y) / P (X)

Bayes -> Naive Bayes

When there are multiple X variables, we simplify it by *assuming the X's are independent*, so the **Bayes** rule

$$P(Y=k \mid X) = \frac{P(X \mid Y=k) * P(Y=k)}{P(X)}$$

where, k is a class of Y

becomes, Naive **Bayes**

$$P(Y=k \mid X1..Xn) = \frac{P(X1 \mid Y=k) * P(X2 \mid Y=k) \ldots * P(Xn \mid Y=k) * P(Y=k)}{P(X1) * P(X2) \ldots * P(Xn)}$$

Probability of Outcome I Evidence (Posterior Probability) = $\frac{\text{Probability of Likelihood of evidence} * \text{Prior}}{\text{Probability of Evidence}}$

Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	Total
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

‘Prior’ probabilities

P(Y=Banana) = 500 / 1000 = 0.50

P(Y=Orange) = 300 / 1000 = 0.30

P(Y=Other) = 200 / 1000 = 0.20

Probability of evidence

P(x1=Long) = 500 / 1000 = 0.50

P(x2=Sweet) = 650 / 1000 = 0.65

P(x3=Yellow) = 800 / 1000 = 0.80

Probability of likelihood of evidences

P(x1=Long | Y=Banana) = 400 / 500 = 0.80

P(x2=Sweet | Y=Banana) = 350 / 500 = 0.70

P(x3=Yellow | Y=Banana) = 450 / 500 = 0.90

Step 4: If a fruit is *'Long', 'Sweet' and 'Yellow'*, what fruit is it?

$$P(\text{Banana} \mid \textit{Long, Sweet and Yellow}) = \frac{P(\text{Long} \mid \text{Banana}) * P(\text{Sweet} \mid \text{Banana}) * P(\text{Yellow} \mid \text{Banana}) \times P(\text{banana})}{P(\text{Long}) * P(\text{Sweet}) * P(\text{Yellow})}$$

$$= \frac{0.8 * 0.7 * 0.9 * 0.5}{P(\text{Evidence})} = 0.252/P(\text{Evidence})$$

P(Orange | *Long, Sweet and Yellow*) = 0, because P(Long | Orange) = 0

P(Other Fruit | *Long, Sweet and Yellow*) = 0.01875 / P(Evidence)

Answer: Banana - Since it has highest probability amongst the 3 classes

Probability of likelihood of evidences

P(x1=Long | Y=Other) = 100 / 200 = 0.5

P(x2=Sweet | Y=Other) = 150 / 200 = 0.75

P(x3=Yellow | Y=Other) = 50 / 200 = 0.25

- The reported accuracy of model was 98.5% but the actual polls results were reported to be predicted with 54.8% accuracy.

- The accuracy of classification of Twitter tweets using sentiment analysis has been questioned by many researchers.

Using Machine learning

USING TWITTER

- The Machine Learning algorithms including NB and SVM were used to classify the tweets.
- The candidate predicted to win the election was based on, Ratio = |P| / |T| where P represents number of positive tweets and T represents the total number of tweets.
- The study reported that using polling data along with sentiments adds value to the predictions at the national level. However, the volume of tweets used for the prediction of electoral results of individual candidates is not an efficient parameter

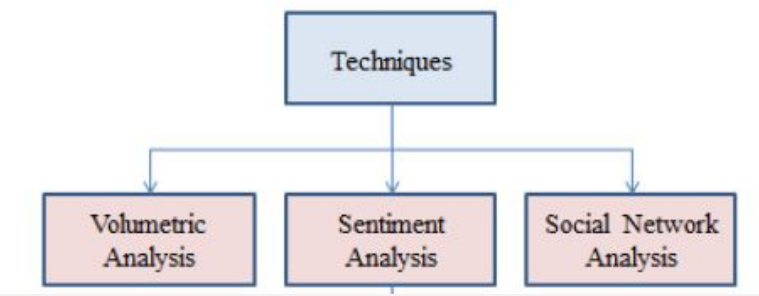
- The effectiveness of social media in predicting electoral results has been investigated for general elections 2013 in Pakistan. Naïve Bayes (NB) was used as classification algorithm.

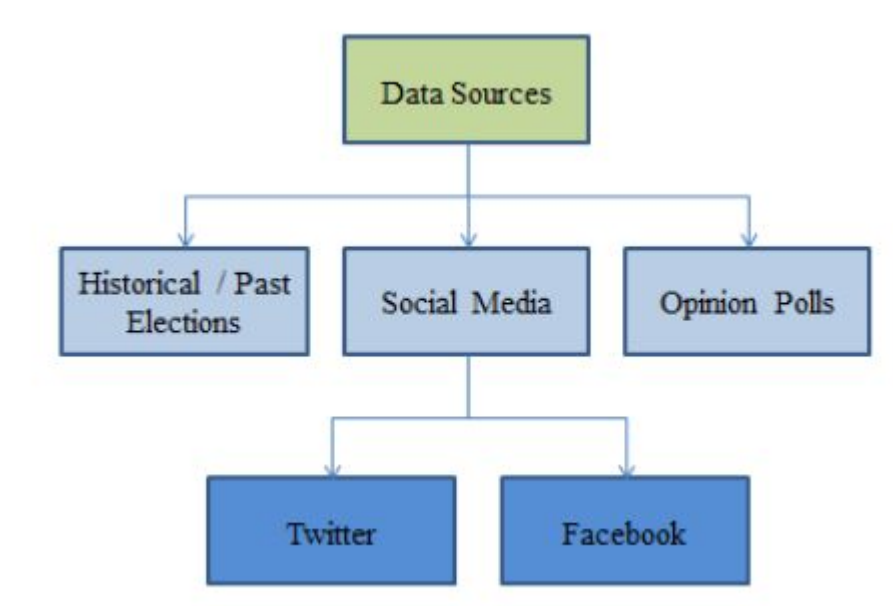
- The accuracy calculated of two only two classes positive and negative, was 70%. But the accuracy eventually drops to 50% when the neutral tweets were included.

- The PTI has 79.29 positive tweets but get the actual polled votes of 20.32%, whereas PMLN has 57.50% positive tweets but get the actual polled votes of 39.35%.

- Hence, Twitter data was reported as a not reliable and non-accurate source for predicting the electoral results in the case for Pakistan general elections 2013.

Summary


Techniques
Volumetric Analysis
Sentiment Analysis
Social Network Analysis


Data Sources
Historical / Past Elections
Social Media
Opinion Polls
Twitter
Facebook

Main Barriers

- Misclassification
- Data Imbalance
- Data Reliability

CONCLUSION

- Majority of studies found the social media data effective in making electoral predictions. Whereas few of the studies also discussed and reported that social media data cannot be relied upon.

Understanding the importance of Data Privacy

References

[1] Muhammad Bilal, Abdullah Gani, Mohsen Marjani, Nadia Malik ; “Predicting Elections : Social media data and techniques” ; 2019 International Conference on Engineering and Emerging Technologies (ICEET) , 2019.

[2] Satish M. Srinivasan, Raghvinder S. Sangwan, Colin J. Neill, Tianhai Zu ; “Twitter data for predicting elections : Insights from emotion classification” ; IEEE Technology and Society Magazine, Vol. 38, Issue : 1, pp: 58 - 63 , 2019.

https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/

“Democracy is not guaranteed, and it is not inevitable. And we cannot let these tech companies have this unchecked power.

We are the ones who have to take back control.”

Carole Cadwalladr

THANKS

It finds out a line/ hyper-plane (in multidimensional space that separate outs classes)