Twitter & Polls: Analyzing and estimating political orientation of Twitter users in India General #Elections2014

Abhishek Bhola
Advisor: Dr. Ponnurangam Kumaraguru

M.Tech Thesis Defense
06-June-2014

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Thesis Committee

- Dr. Ayesha Choudhary, JNU
- Dr. Vinayak Naik, IIIT-Delhi
- Dr. PK (Chair), IIIT-Delhi

Presentation Outline

- Research Motivation
- Research Aim
- Related Work
- Research Gap
- Data Analysis
- Classification of Political Orientation
- System Design
- Conclusion
- Limitations and Future Work

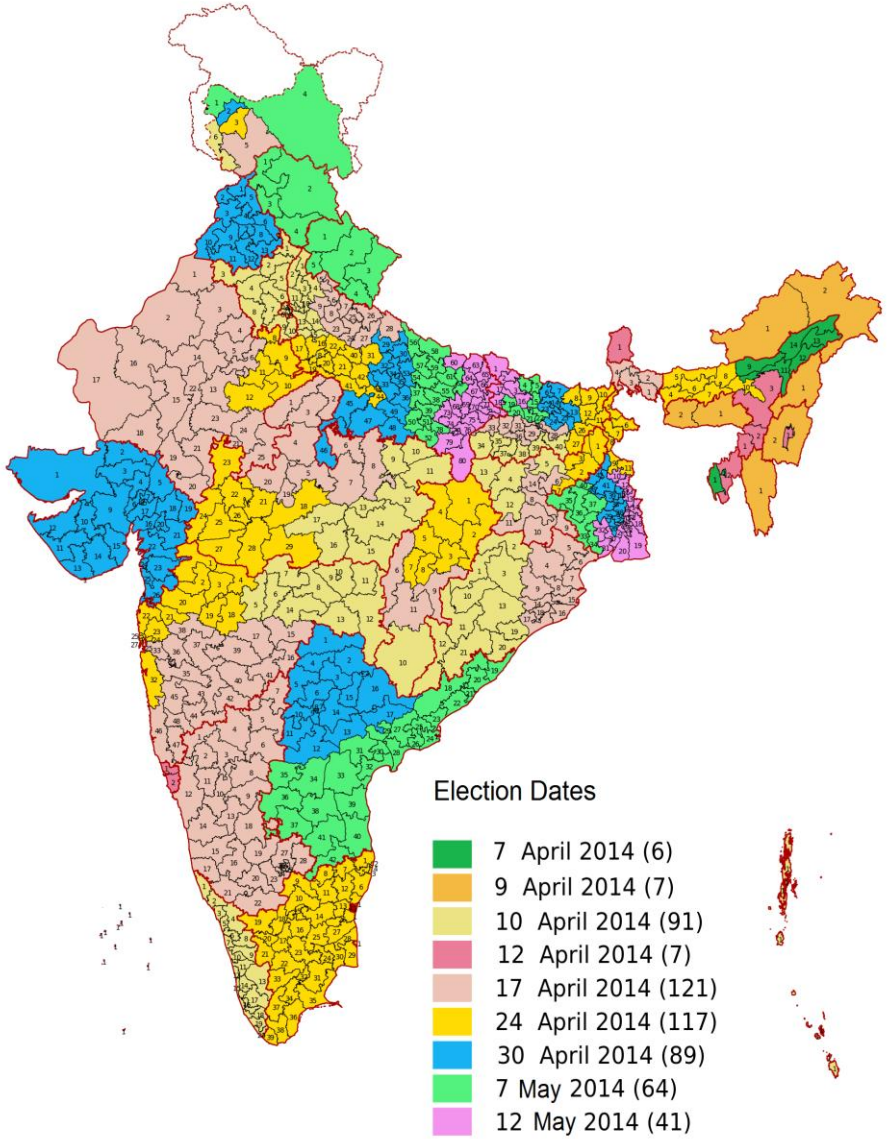
Presentation Outline

- Research Motivation
- Research Aim
- Related Work
- Research Gap
- Data Analysis
- Classification of Political Orientation
- System Design
- Conclusion
- Limitations and Future Work

India General Elections 2014

- Elections for the 16th Lok Sabha
- Total Seats: 543
- Number of Registered Voters: 813 million
- Newly registered voters: 100 million
- Money at stake: $5 billion
- Number of parties registered with the Election Commission: 1616
- National Parties: 6
- State Parties: 47
- Number of candidates: 8000

India General Elections 2014



Elections and Social Media

- Major parties battling it out:
  - Aam Aadmi Party (AAP)
  - Bhartiya Janta Party (BJP)
  - Indian National Congress (INC)
- Their PM Candidates:
  - Arvind Kejriwal
  - Narendra Modi
  - Rahul Gandhi
- Internet Users in India: 243 million (by June 2014)
- Facebook Users: 114.8 million
- Twitter Users: 33 million

Elections and Social Media

- Google+ Hangout: Interaction with party workers
- What'sApp: To send bulk messages
- Facebook: Televised Interviews and ad campaigns
- Instagram: Pictures of party rallies uploaded
- YouTube: Videos of rallies uploaded
- Google's Election Hub

Research Motivation

- Social Media could sway 3-4% of urban votes: IAMAI
- Increase in elections related data- 600% from 2009
- Almost all leaders and parties are on Twitter
- Extensively used for communicating and interaction

Presentation Outline

- Research Motivation
- Research Aim
- Related Work
- Research Gap
- Data Analysis
- Classification of Political Orientation
- System Design
- Conclusion
- Limitations and Future Work

Research Aim

- To analyze and draw meaningful inferences from the collection of tweets collected over the entire duration of elections
- To check the feasibility of development of a classification model to identify the political orientation of the twitter users based on the tweet content and other user based features.
- To develop a system to analyze and monitor the election related tweets on daily basis.

Presentation Outline

- Research Motivation
- Research Aim
- Related Work
- Research Gap
- Data Analysis
- Classification of Political Orientation
- System Design
- Conclusion
- Limitations and Future Work

Related Work

1

Tumasjan et. al. *Predicting elections with twitter: What 140 characters reveal about political sentiment*, 2010

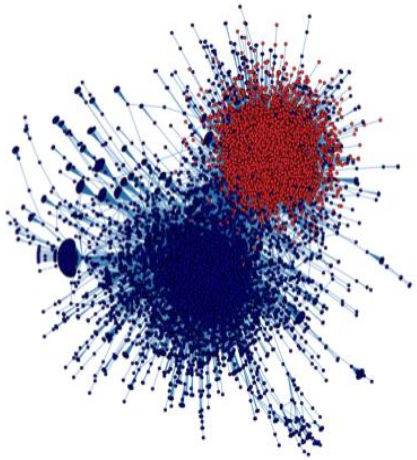
2

Jungherr et. al. *Why the pirate party won the german election of 2009 or the trouble with predictions*, 2012

3

Skoric et. al. *Tweets and votes: A study of the 2011 Singapore general election*, 2012

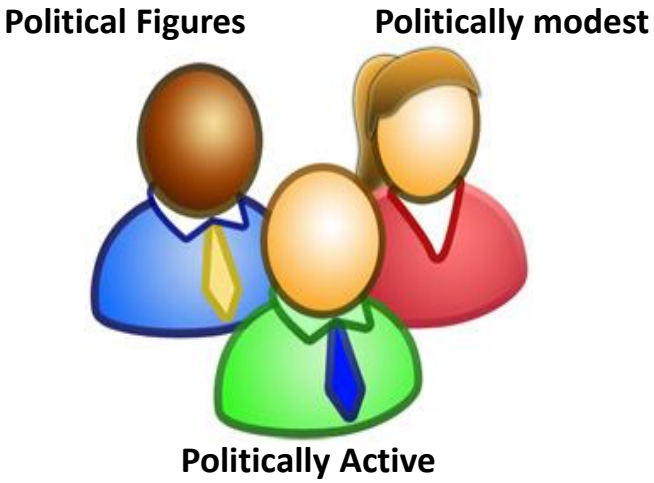
4



Conover et. al. *Predicting the political alignment of twitter users*, 2011

Related Work

5



Cohen et. al.
Classifying political orientation on twitter: Its not easy!, 2013

6

Simplify360:
Calculation of SSI

7

NExT Centre, NUS
Weekly Infographics

8

Twitritis +
Wright State
University

Presentation Outline

- Research Motivation
- Research Aim
- Related Work
- Research Gap
- Data Analysis
- Classification of Political Orientation
- System Design
- Conclusion
- Limitations and Future Work

Research Gap

- No previous attempts to classify the political orientation of users in the Indian scenario
- No previous work explored both ‘Pro’ as well as ‘Anti’ views

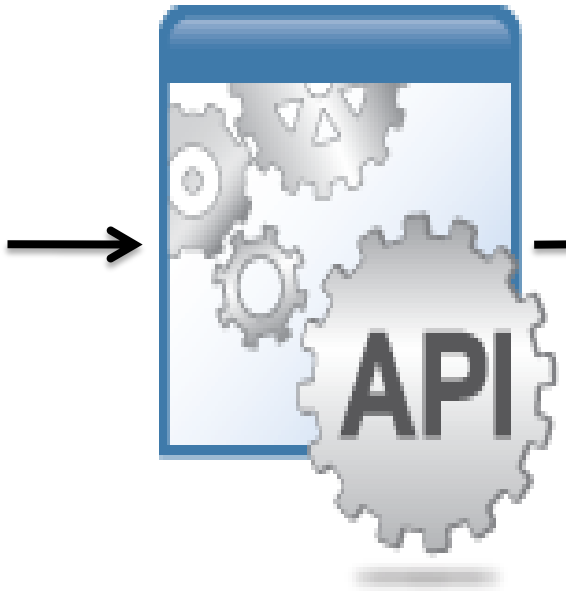
Presentation Outline

- Research Motivation
- Research Aim
- Related Work
- Research Gap
- Data Analysis
- Classification of Political Orientation
- System Design
- Conclusion
- Limitations and Future Work

Data Collection

Keywords for Election related tweets

Streaming



Elections related tweets

Data Collection

Screen names of Twitter Profiles

Streaming

130 Twitter Profiles

Twitter Data

As per reports, what Twitter had:

- #Tweets **till April 30**: **49 million***
- #Tweets, **Jan 1- May 12**: **56 million**ǂ
- 600% rise in #Tweets from 2009 elections
- 2009 elections only 1 politician had an account with 6K followers

What we had with us:

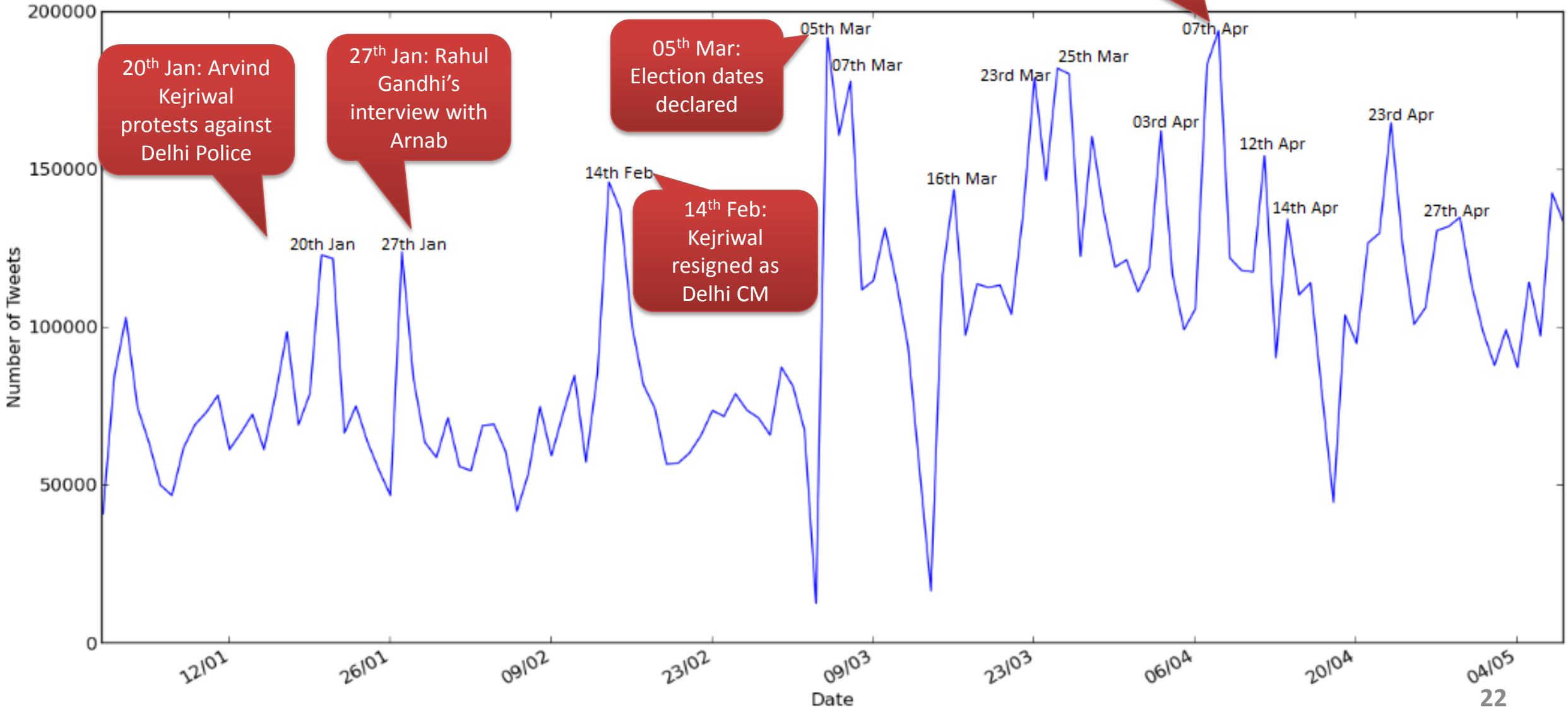
- #Tweets, **Sept 25 – May 16**: **18.21 million**
- #Tweets, **Jan 1- May 12**: **13.09 million**
- 23.37 % of Tweets
- Difference in the keywords used

ǂ Twitter's official blog: https://blog.twitter.com/2014/indias-2014-twitterelection
* India Today Post: http://bit.ly/1hukpDE

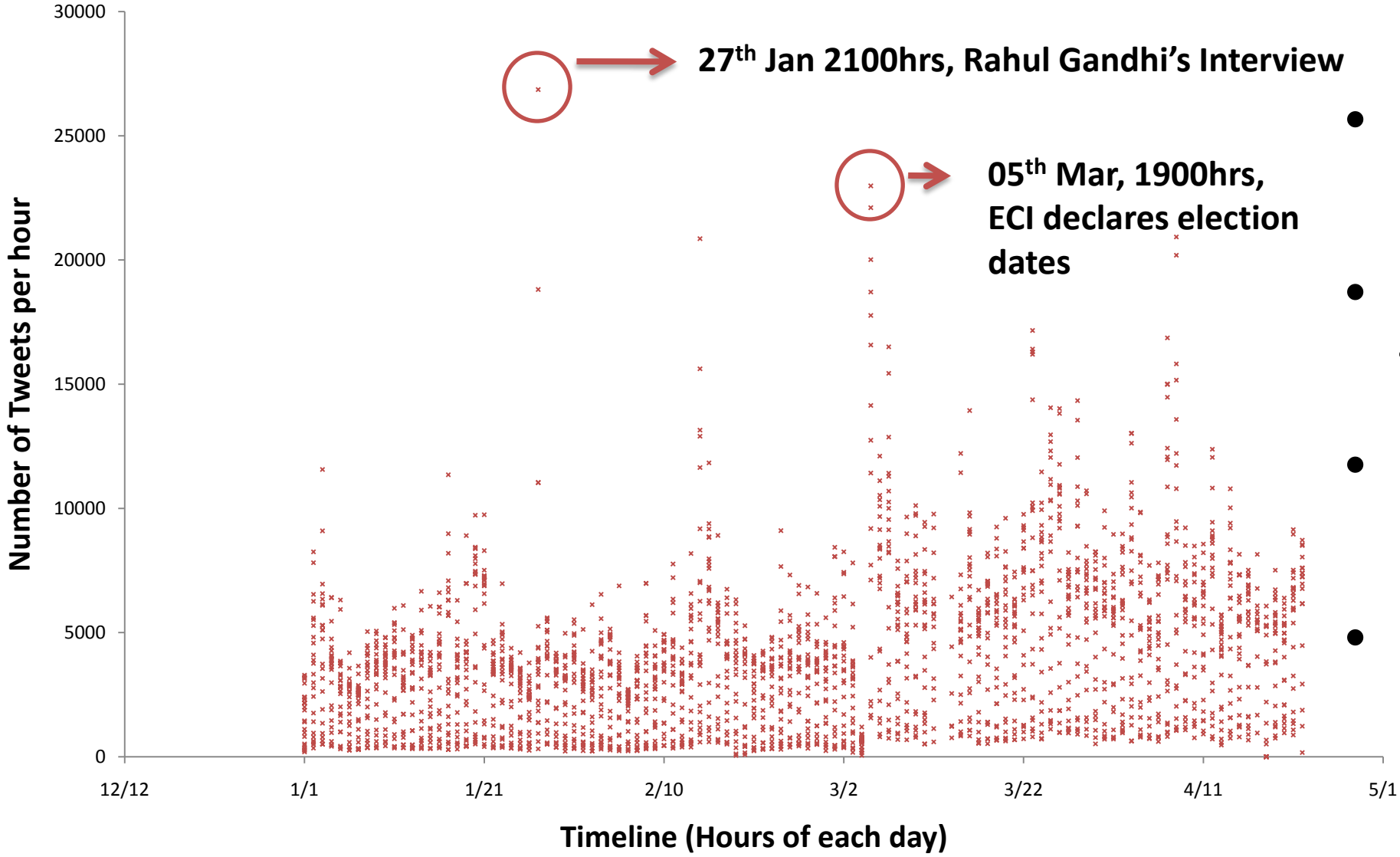
Methodology for Analysis

- Tweets were picked from the database
- Different fields were exploited for different analysis
- Python, Matplotlib and Excel were used to plot graphs

Volume Analysis

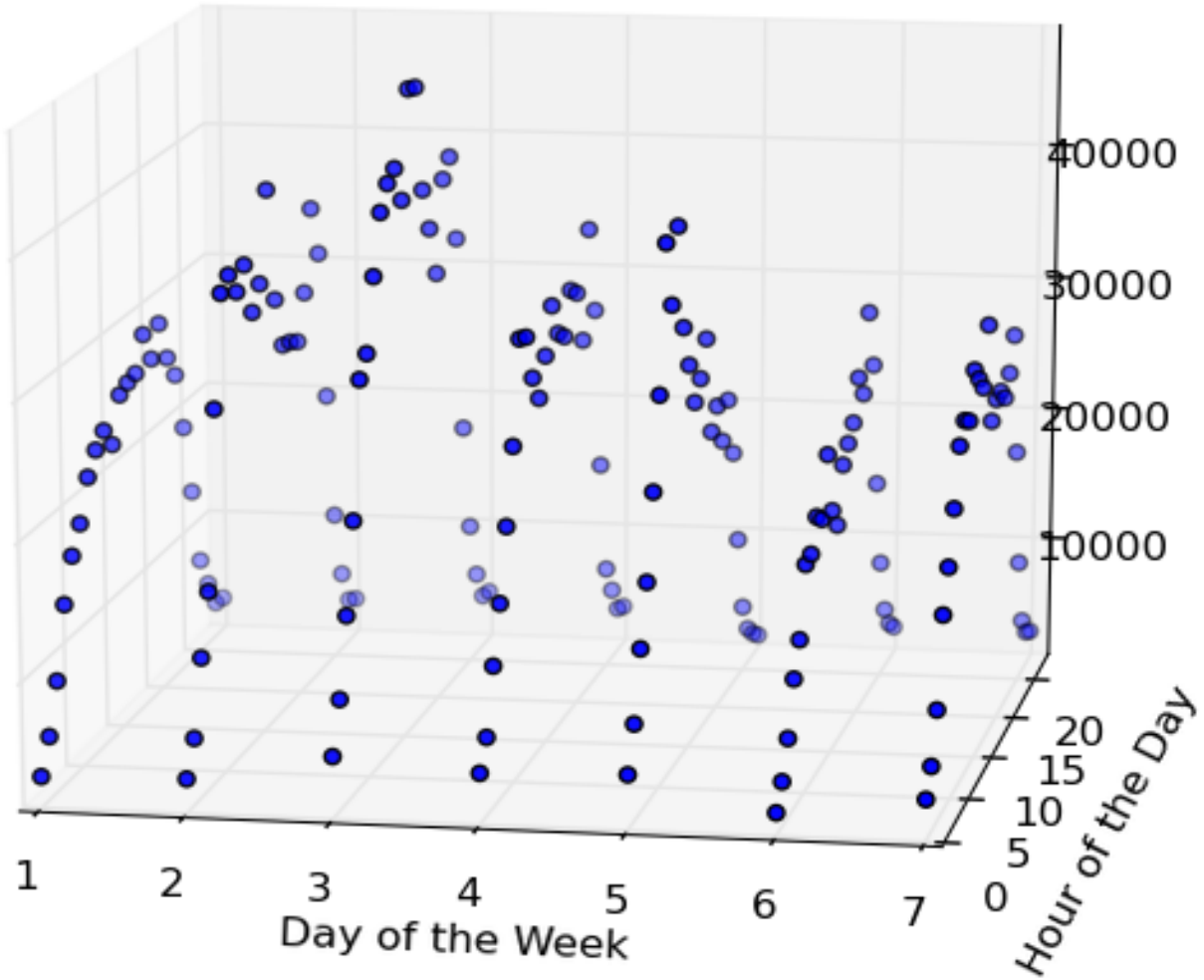


Hourly Frequency Analysis



- Per hour tweets frequency
- Mean: 4073.08 Tweets
- Std deviation: 3072.77
- UCL: 10218.63

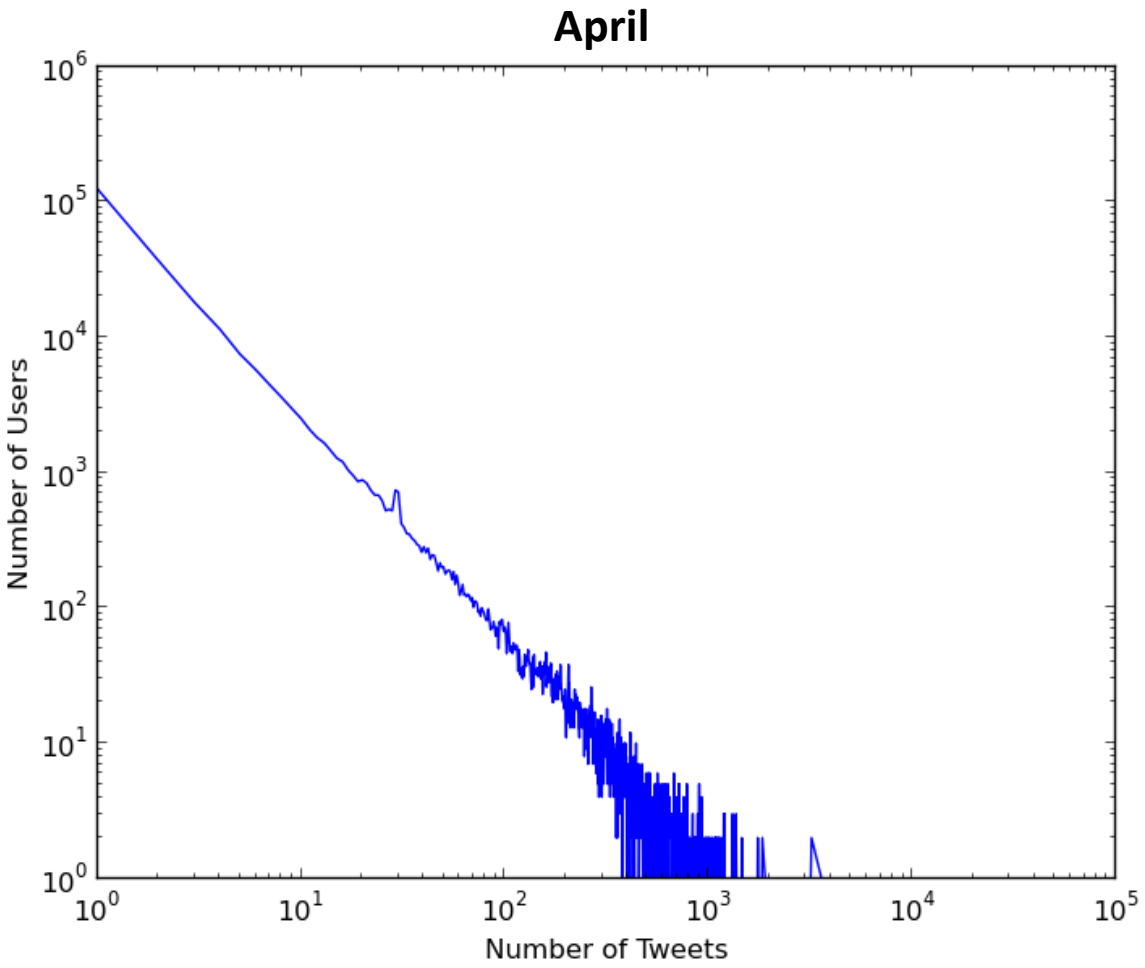
Hour v/s Day of the Week



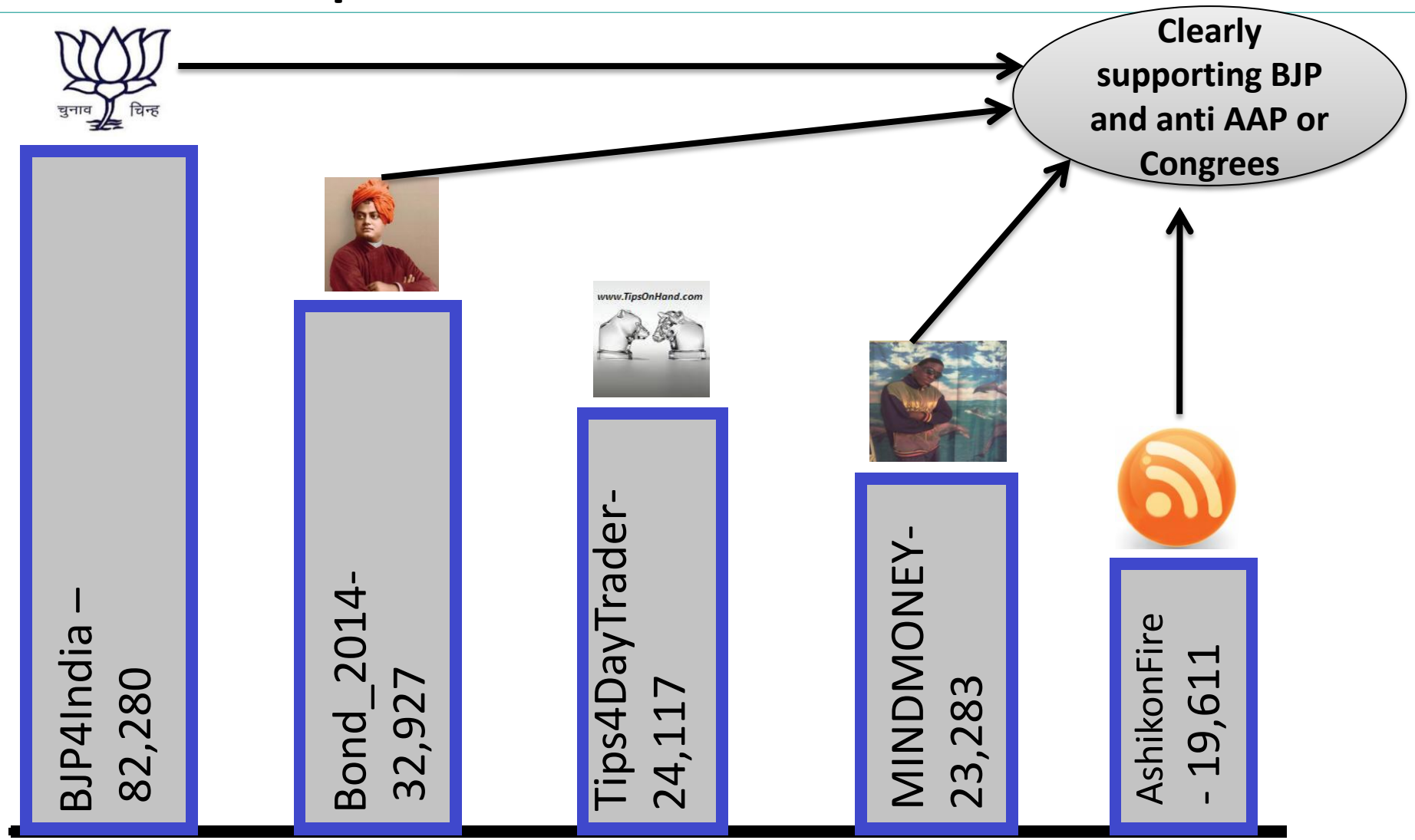
- Max #Tweets during weekdays
- Max Tweets on Tuesdays and Wednesdays
- Tweeting activity goes higher during the second half of the day
- Most of the relevant events were on weekdays

Who tweets how much

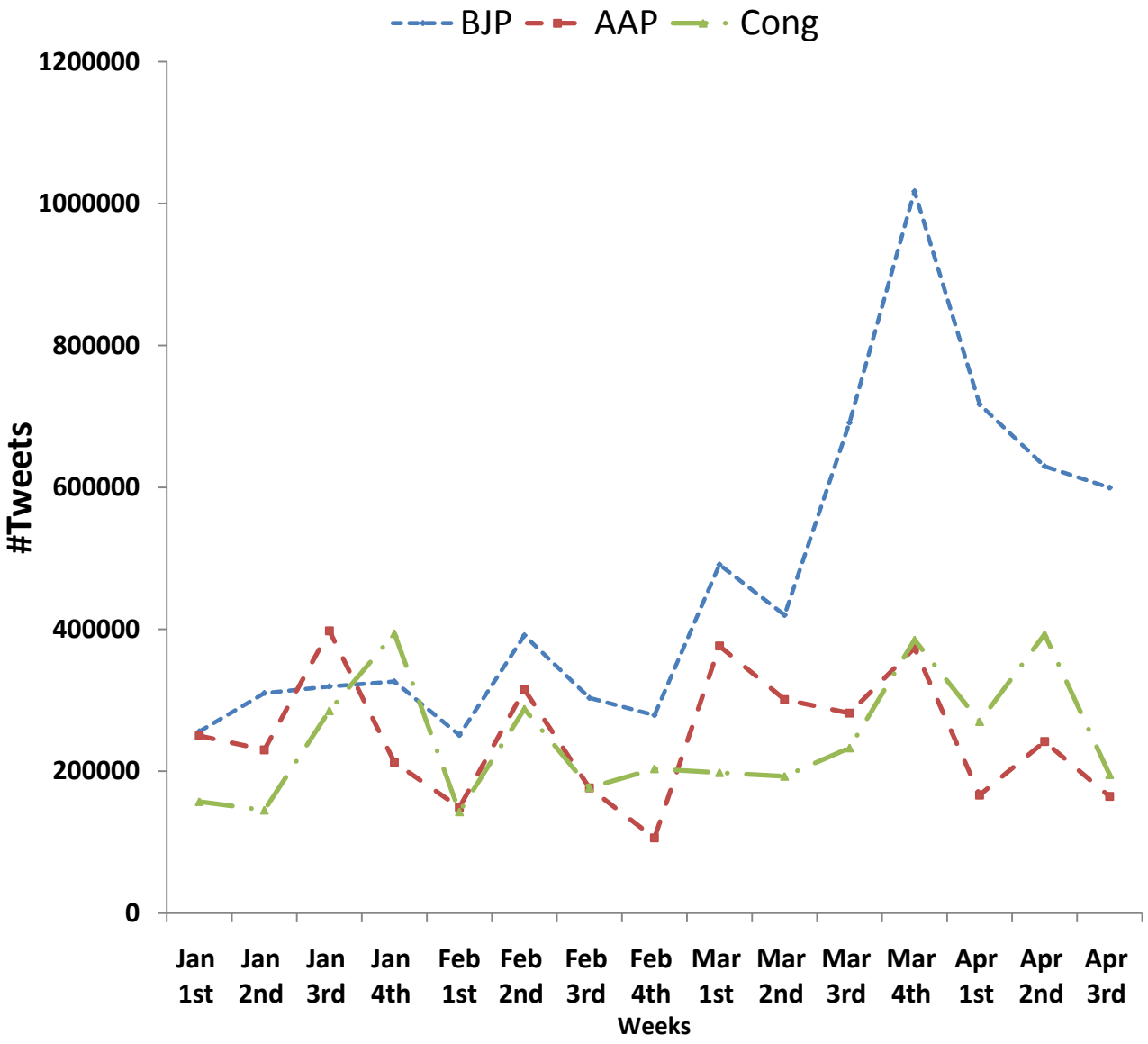
- Inverse Law Proportion
- Graph for the tweets of month of all the months
- April has the highest number of unique users
- April is the month with a single user tweeting >10^4 times
- **815,425** total unique users

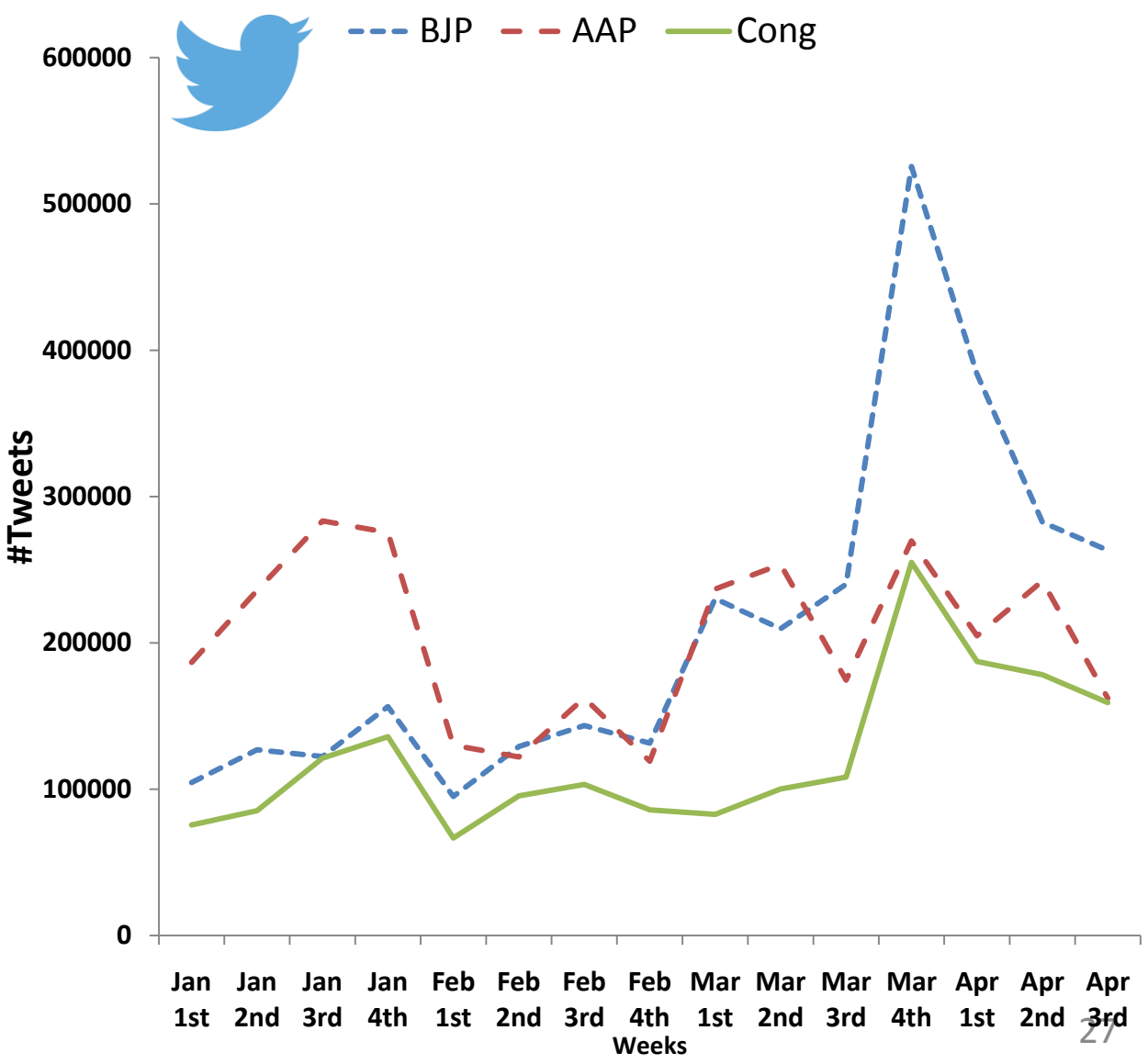


Top 5 active tweeters

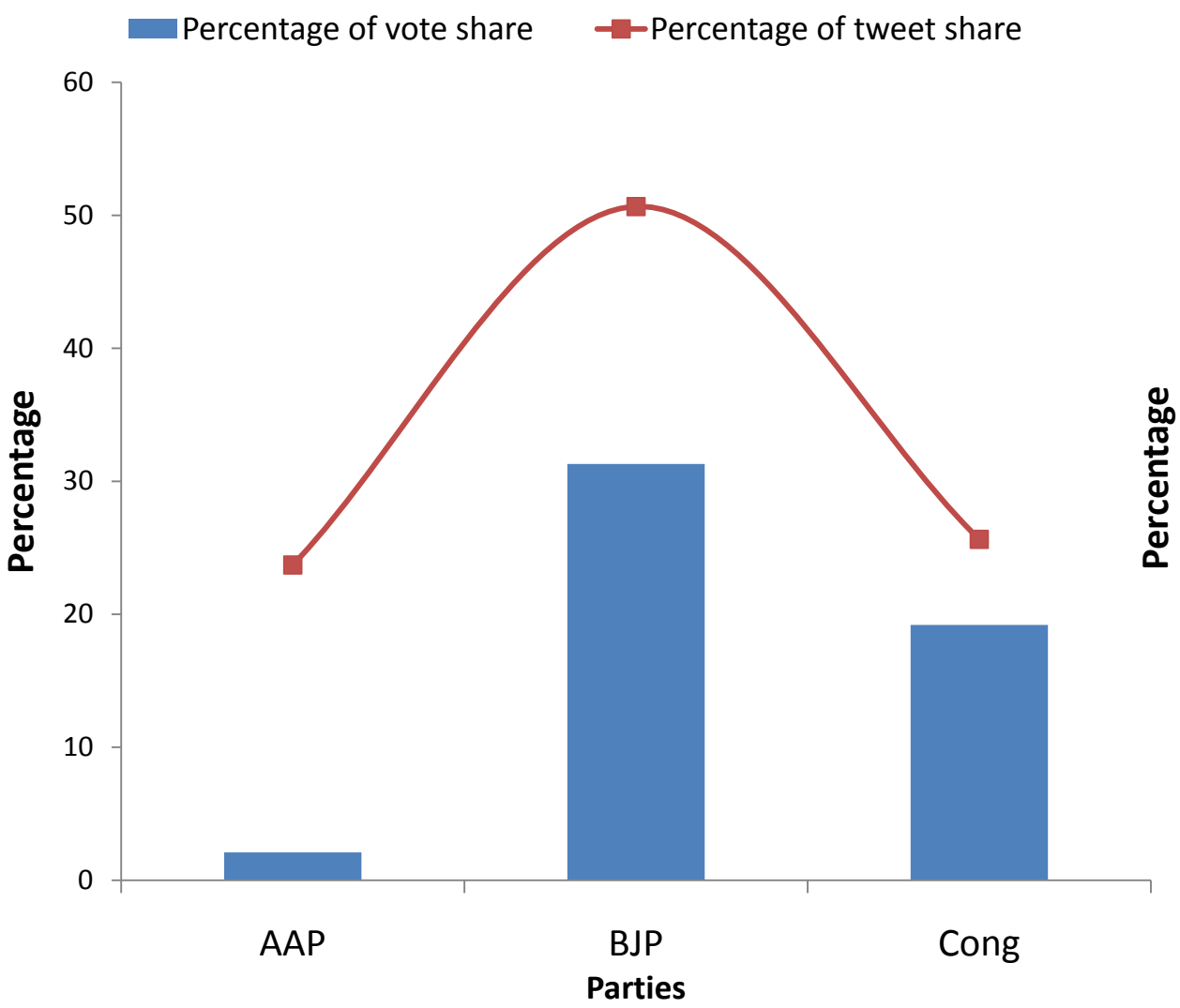


Which party received maximum mentions

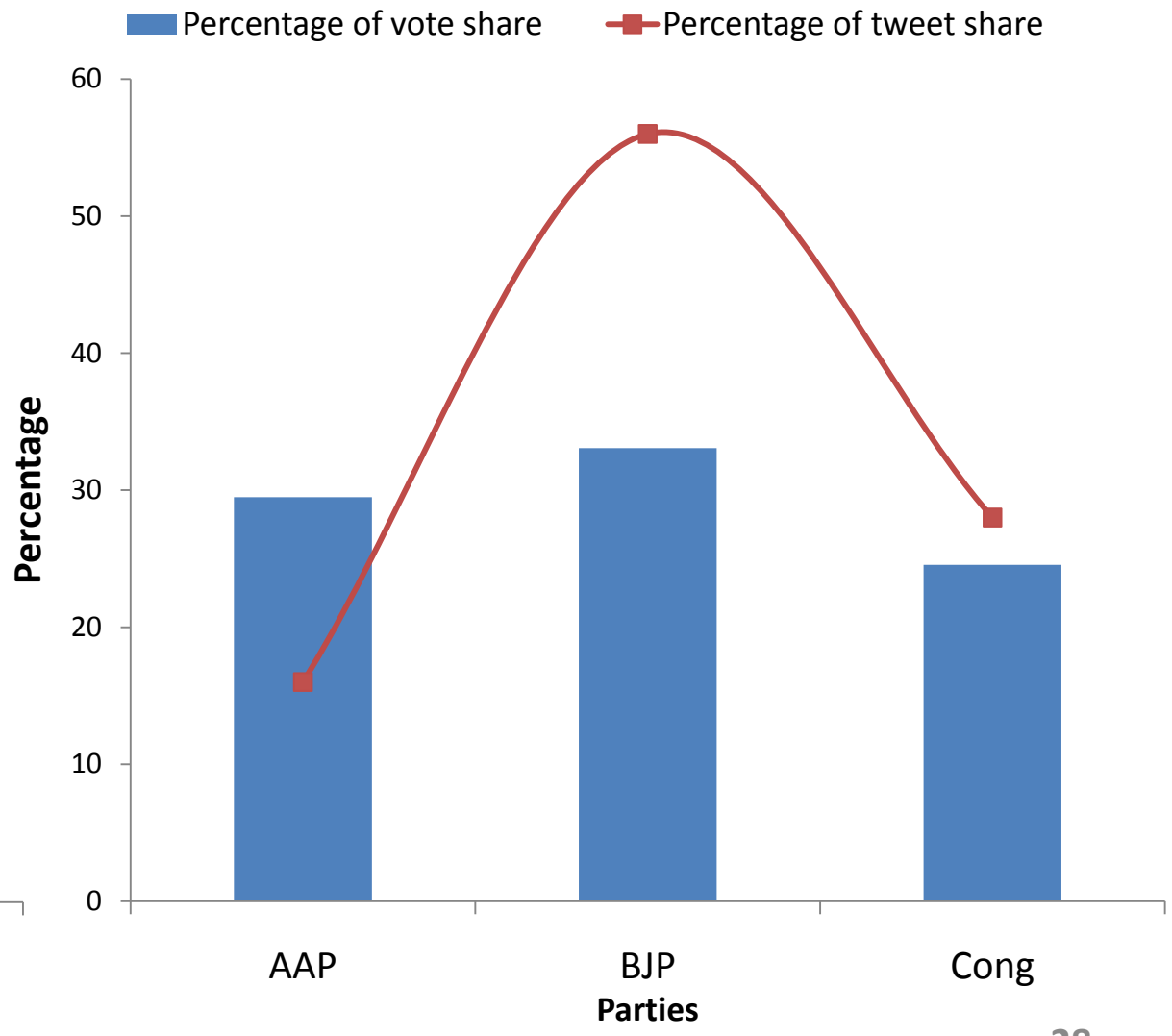




Comparing electoral results with tweet share

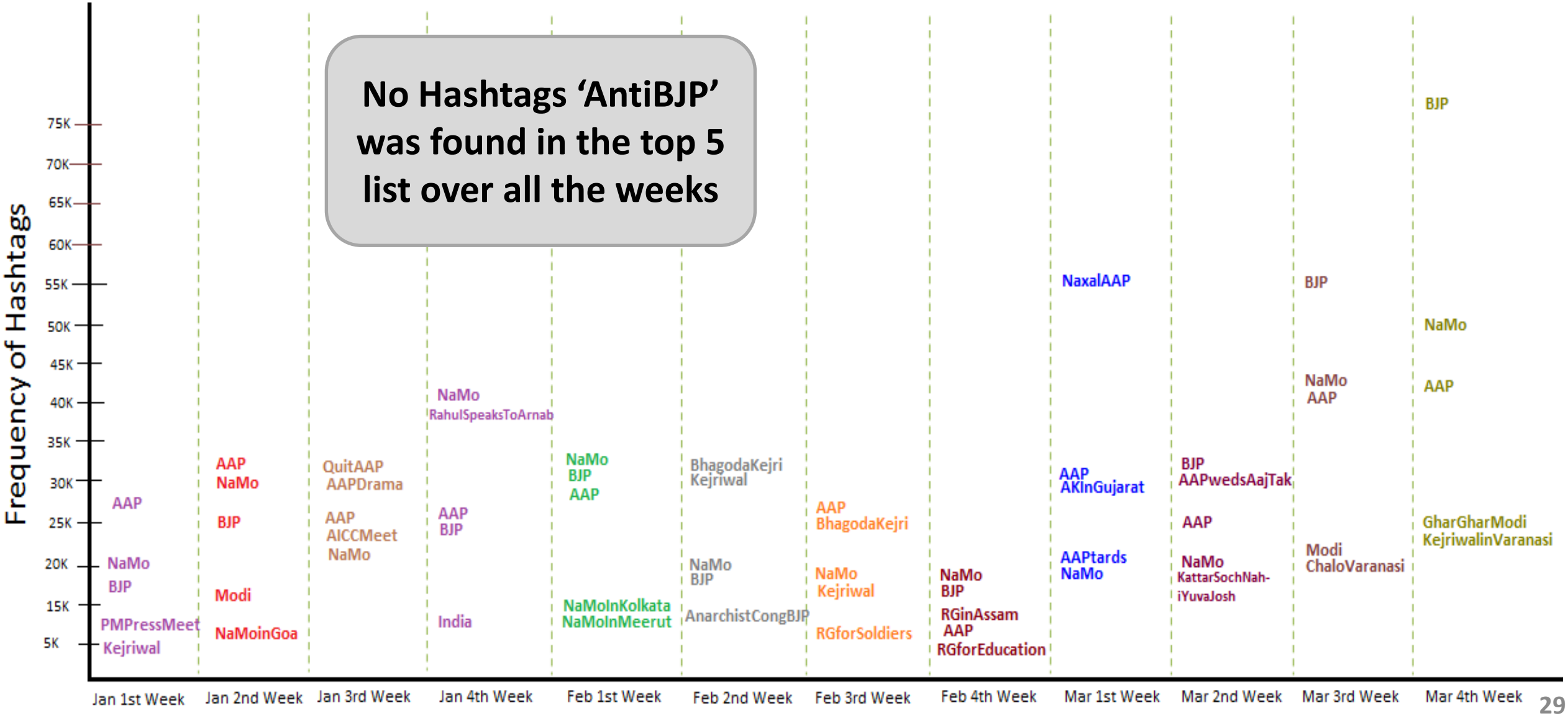


General Elections 2014

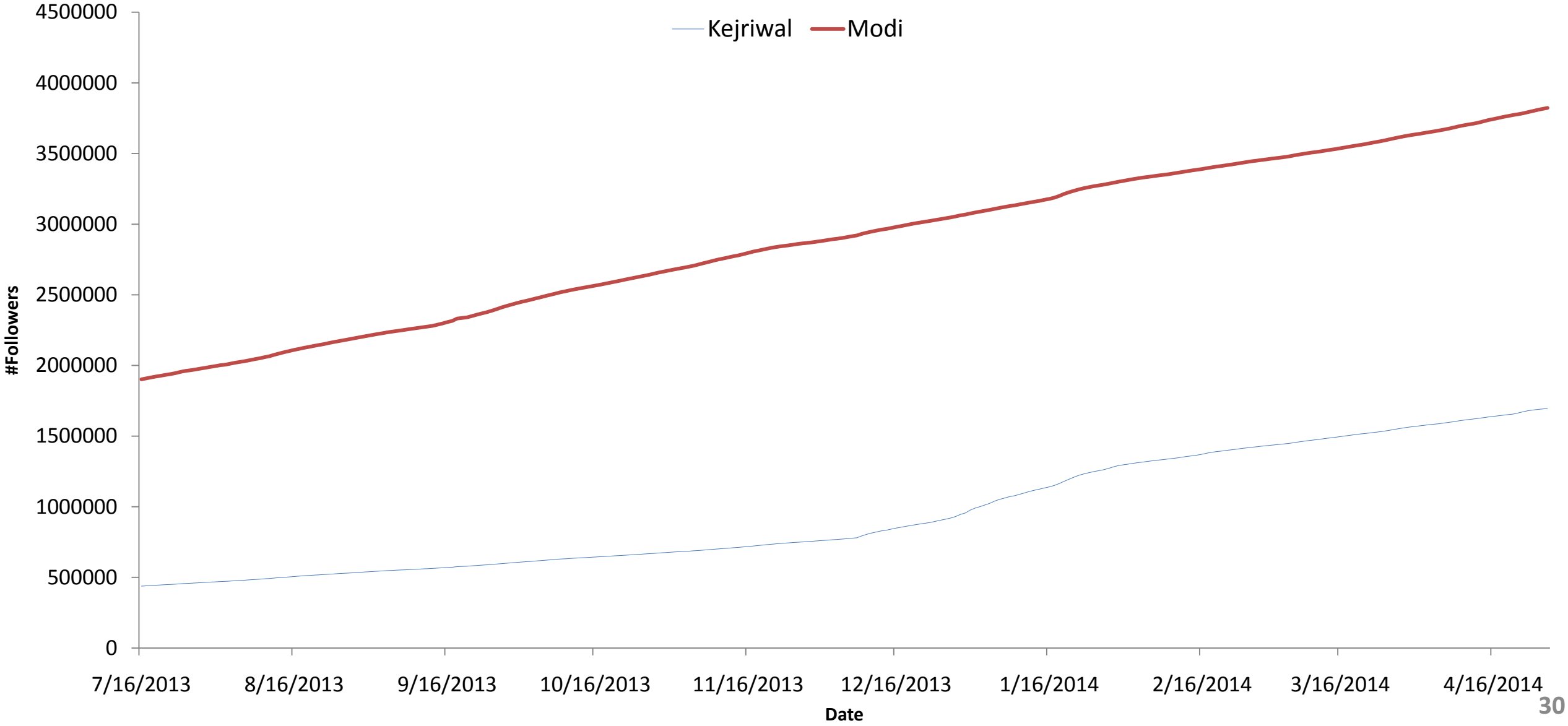


Delhi Assembly Elections 2013

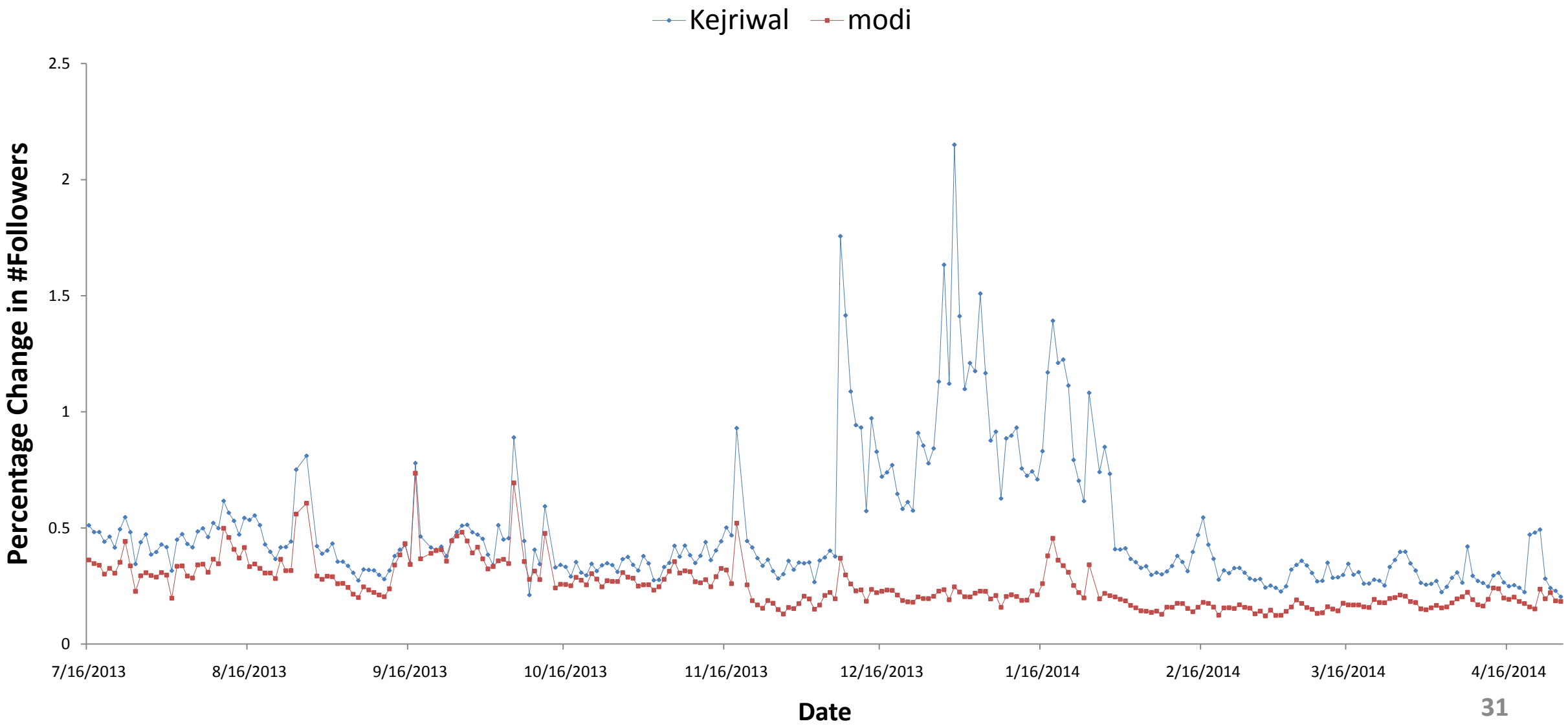
Top 5 Hashtags of every week



Analyzing the popularity of Modi & Kejriwal: #Followers

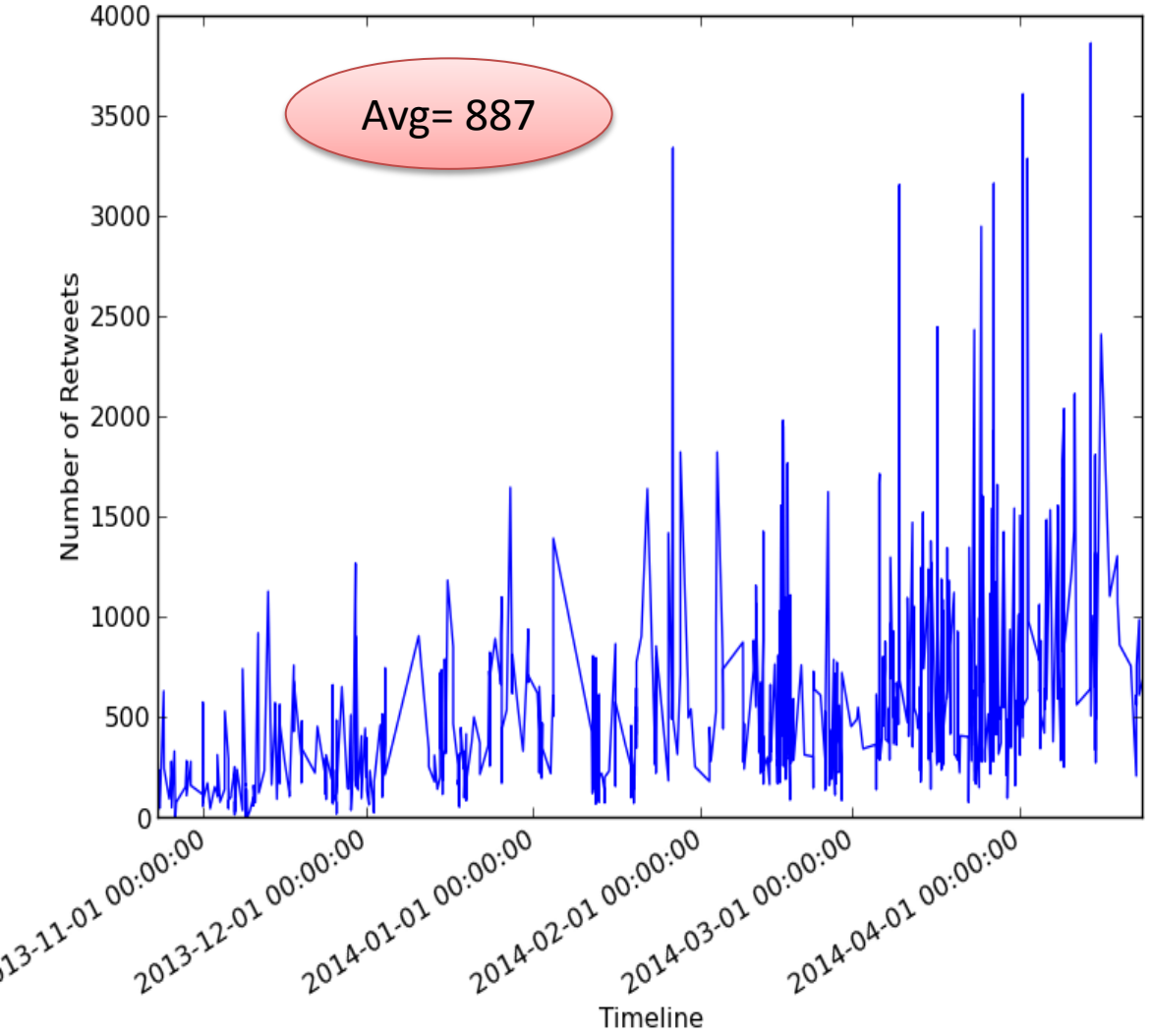


Analyzing the popularity of Modi & Kejriwal: #Followers

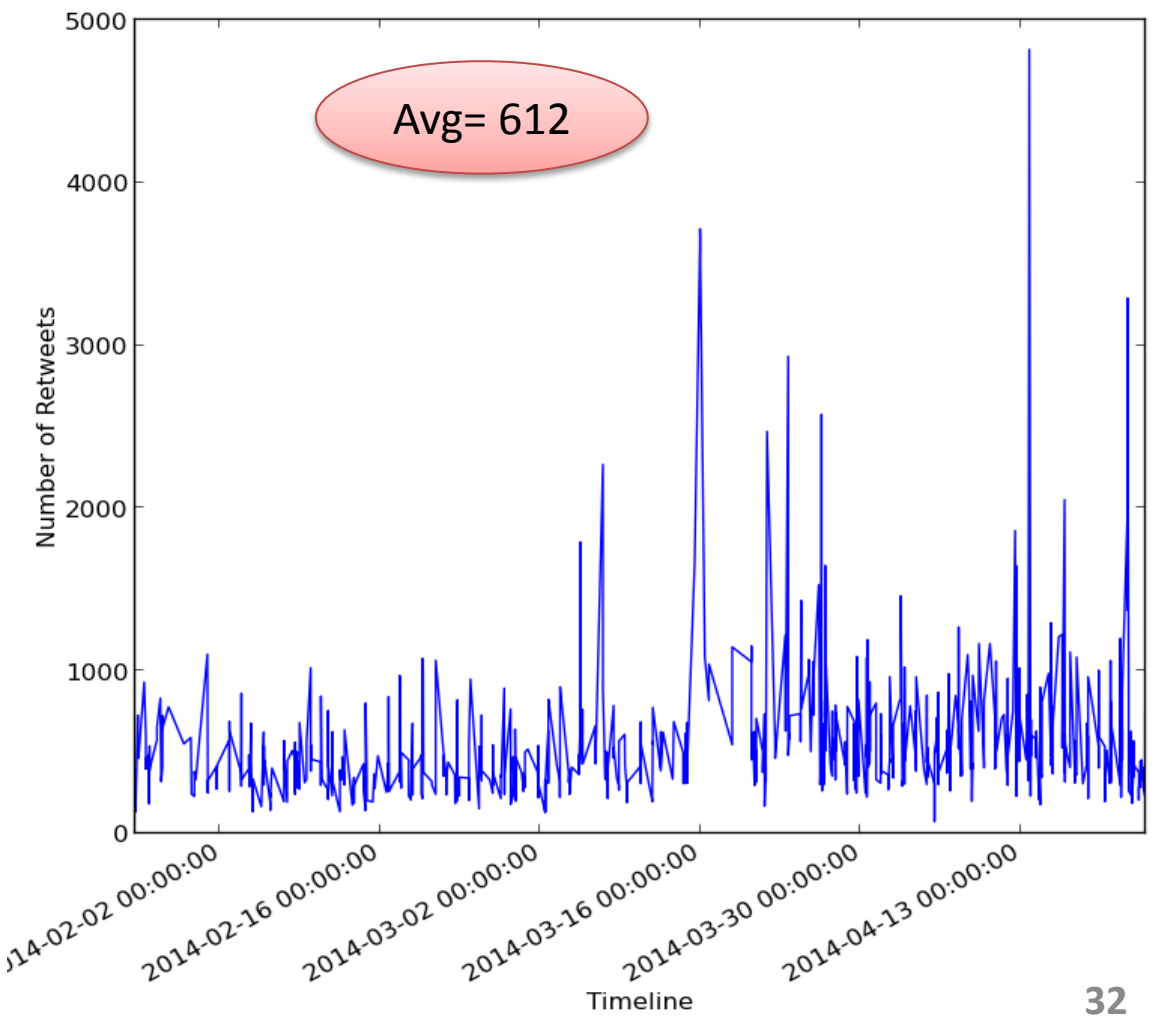


Analyzing the popularity of Modi & Kejriwal: Retweet Frequency on Tweets

Kejriwal

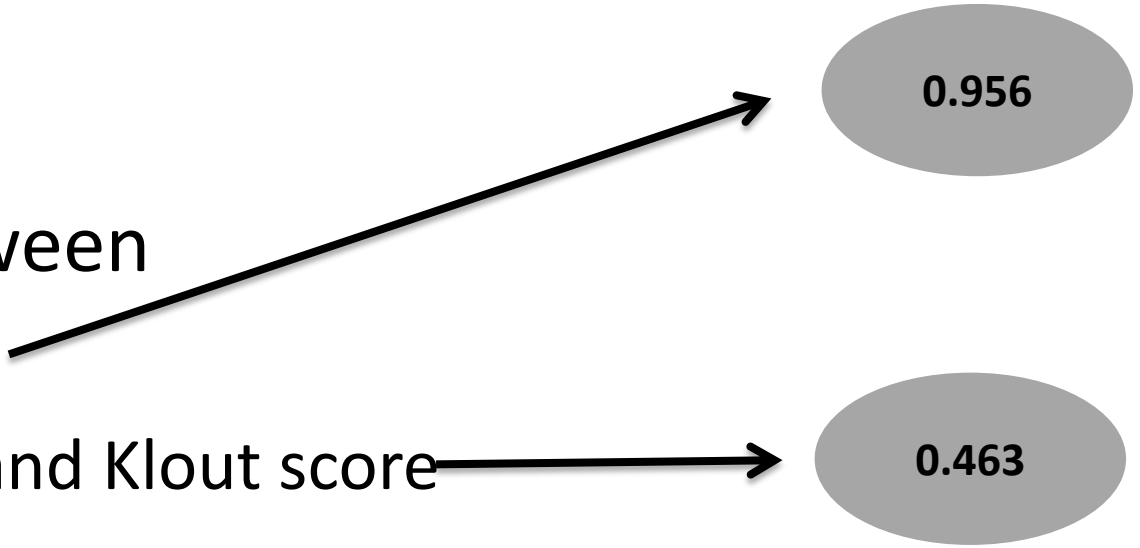


Modi



Tradeoff b/w #Followers & Retweet Frequency

- Klout score uses a lot of factors viz.,
  - #Followers
  - #Friends
  - #Retweets on each tweet
- Pearson's correlation between
  - #Followers and Klout score
  - Avg. #retweets on tweets and Klout score



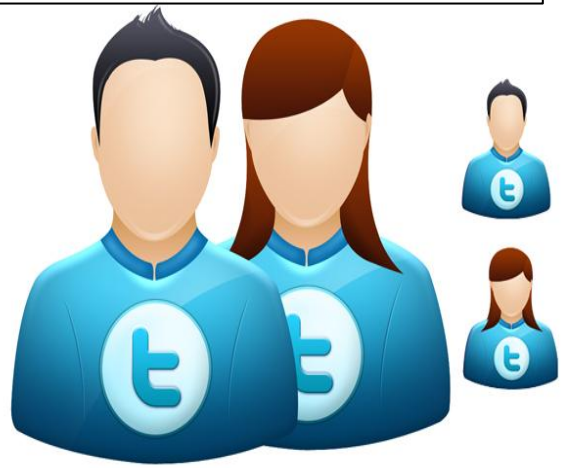
Narendra Modi's popularity was more than Arvind Kejriwal based on the number of followers

Presentation Outline

- Research Motivation
- Research Aim
- Related Work
- Research Gap
- Data Analysis
- Classification of Political Orientation
- System Design
- Conclusion
- Limitations and Future Work

Political Orientation of Users

1000 random twitter user profiles tweeting about India Elections were selected



10 Data Annotators decided their political orientation on the basis of tweets

Pro
- AAP
- BJP
- Cong
- Can't Say

Anti
- AAP
- BJP
- Cong
- Can't Say

Other information about these profiles such as #Followers, #Friends, Tweets between Mar 20 – Apr 10 was also collected

REST v1.1.

Agreement between the annotators

- Confusion Matrix for the 1st set of 250 instances (Pro)

Annotator 1

Annotator 2	AAP	BJP	CONG	CAN'T SAY
AAP	18	4	0	3
BJP	6	76	1	21
CONG	0	4	2	3
CAN'T SAY	11	11	4	86

Observed agreement:
183/250 = **0.732**

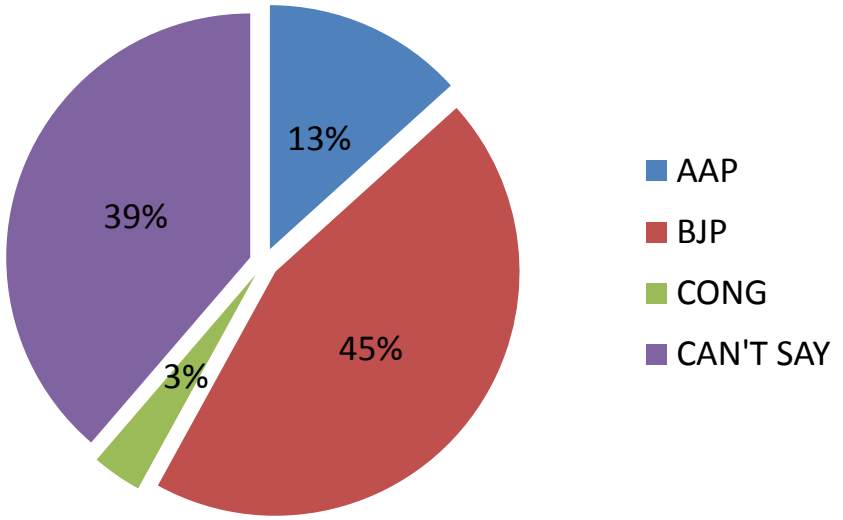
Cohen's Kappa coefficient, $\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$

Hypothetical Chance agreement

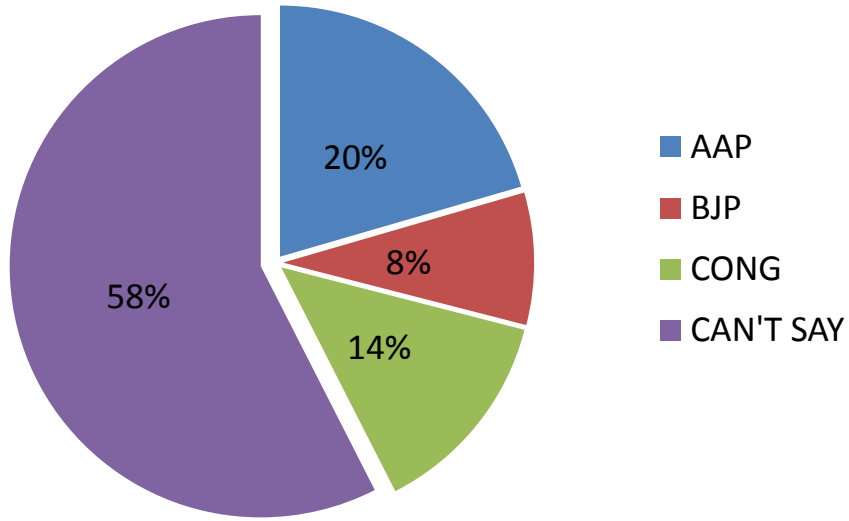
- *Pr(e)* = 0.375
- κ = 0.571

Annotation Results

Party	Pro	Anti
AAP	133	205
BJP	447	85
CONG	33	135
CAN’T SAY	387	575



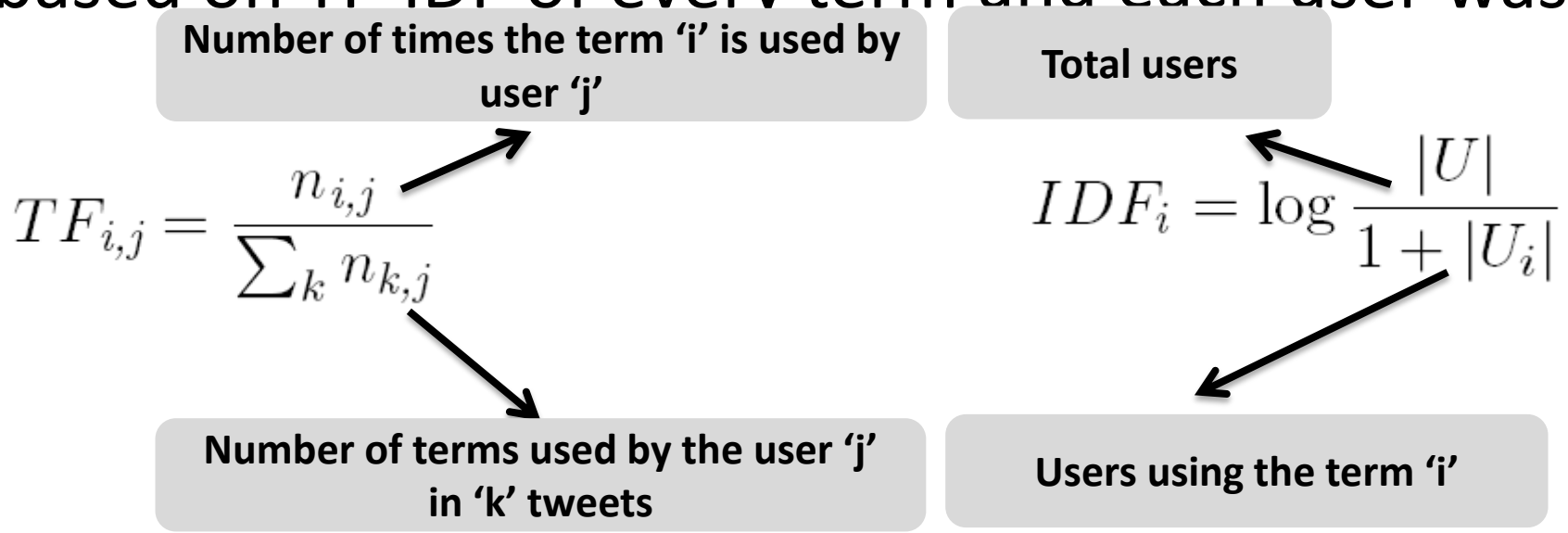
PRO



ANTI

Text Based Classification

- Text of 200 tweets collected
- Stop words, URLs, Hashtags and user mentions removed
- Words like ‘RT’, ‘&amp’, ‘ka’, ‘ke’, ‘ki’ etc. were also removed
- Vector based on TF-IDF of every term and each user was formed



Text Based Classification: Results

*Results for all 613 '**Pro**' Instances*

Instances: 613			
Attributes: 9312			
Classifier: Random Forest, 10 folds cross-validation			
Efficiency: 72.36%			
Party	Precision	Recall	F-measure
AAP	0.381	0.061	0.105
BJP	0.736	0.975	0.839
CONG	0	0	0

Good Efficiency

But 0 Precision and Recall for Congress

*Results for equal '**Pro**' Instances*

Instances: 99			
Attributes: 2442			
Classifier: Random Forest, 10 folds cross-validation			
Efficiency: 42.42%			
Party	Precision	Recall	F-measure
AAP	0.424	0.758	0.543
BJP	0.481	0.394	0.433
CONG	0.308	0.121	0.174

Efficiency falls

Precision and Recall for Congress not 0

- We tried 2-class classification with all 3 possible pairs and got 65.15% efficiency for AAP-BJP

Text Based Classification: Results

*Results for all 425 '**Anti**' Instances*

Instances: 425			
Attributes: 8014			
Classifier: Random Forest			
Efficiency: 47.75%			
Party	**Pre**	**Recall**	**F-measure**
AAP	0.489	0.863	0.624
BJP	0.313	0.059	0.099
CONG	0.447	0.157	0.232

Less instances of BJP result in low Precision and recall values

*Results for equal '**Anti**' Instances*

Instances: 255			
Attributes: 6847			
Classifier: Random Forest			
Efficiency: 37.25%			
Party	**Pre**	**Recall**	**F-measure**
AAP	0.321	0.529	0.4
BJP	0.47	0.365	0.411
CONG	0.388	0.224	0.284

Hashtags Based Classification

- Hashtags represent the topic of the tweet
- Picked up all the hashtags in the last 200 tweets of the user
- Computed the user vector in same manner as in text based classification
- Terms in this case were the hashtags instead of words used in the tweets

Hashtags Based Classification: Results

*Results for all 613 '**Pro**' Instances*

Instances: 613			
Attributes: 1398			
Classifier: Random Forest			
Efficiency: 75.49%			
Party	**Pre**	**Recall**	**F-measure**
AAP	0.759	0.167	0.273
BJP	0.756	0.983	0.856
CONG	0	0	0

*Results for all 425 '**Anti**' Instances*

Instances: 425			
Attributes: 1182			
Classifier: Random Forest			
Efficiency: 50.35%			
Party	**Pre**	**Recall**	**F-measure**
AAP	0.5	0.946	0.654
BJP	0.875	0.165	0.277
CONG	0.286	0.045	0.077

- Efficiency improved by 2-3%, even with equal number of instances
- Precision and recall values remain 0 for Congress with all 'Pro' instances

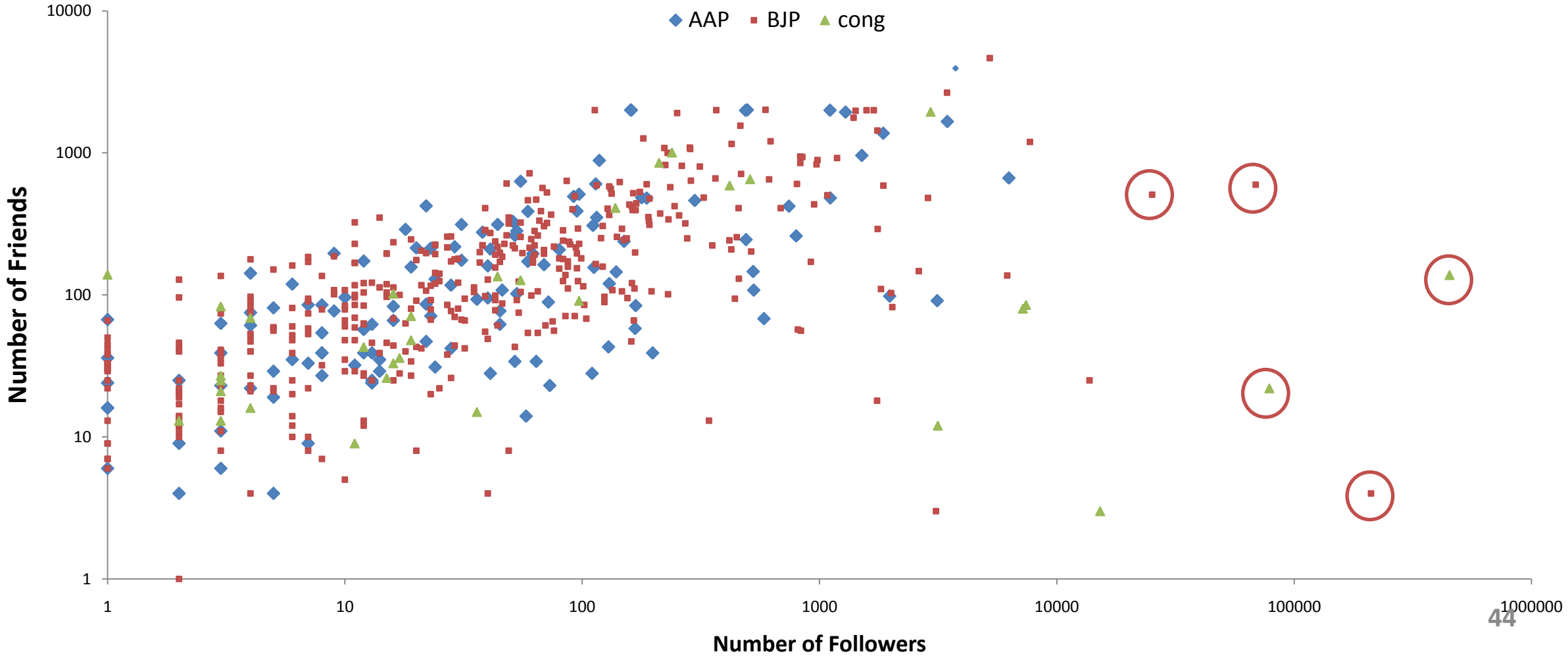
User Features Based Classification

1. #Friends
2. #Followers
3. Following AAP?
4. Following BJP?
5. Following Congress?
6. #AAP related words
7. #BJP related words
8. #Congress related words
9. #AAP related hashtags
10. #BJP related hashtags
11. #Congress related hashtags

User Features Based Classification

1. #Friends
2. #Followers

Scatter Plot b/w #Followers and #Friends for users 'Pro' to each party



User Features Based Classification: Results

*Results for all 613 '**Pro**' Instances*

Instances: 613			
Attributes: 12			
Classifier: Random Forest			
Efficiency: 71.45%			
Party	Pre	Recall	F-measure
AAP	0.455	0.376	0.412
BJP	0.781	0.855	0.816
CONG	0.429	0.182	0.255

Good Efficiency

Precision and Recall not 0

*Results for equal '**Pro**' Instances*

Instances: 99			
Attributes: 12			
Classifier: Random Forest			
Efficiency: 50%			
Party	Pre	Recall	F-measure
AAP	0.483	0.412	0.44
BJP	0.439	0.486	0.462
CONG	0.588	0.606	0.597

Improvement in efficiency by 6.4%

User Features Based Classification: Results

*Equal instances of '**Pro**' AAP-BJP*

Instances: 266			
Attributes: 9			
Classifier: Random Forest			
Efficiency: 61.27%			
Party	**Pre**	**Recall**	**F-measure**
AAP	0.606	0.62	0.615
BJP	0.781	0.59	0.603

2-Class Classification

*Equal instances of '**Pro**' AAP-Cong*

Instances: 66			
Attributes: 9			
Classifier: Random Forest			
Efficiency: 71.21%			
Party	**Pre**	**Recall**	**F-measure**
AAP	0.706	0.727	0.716
CONG	0.719	0.697	0.708

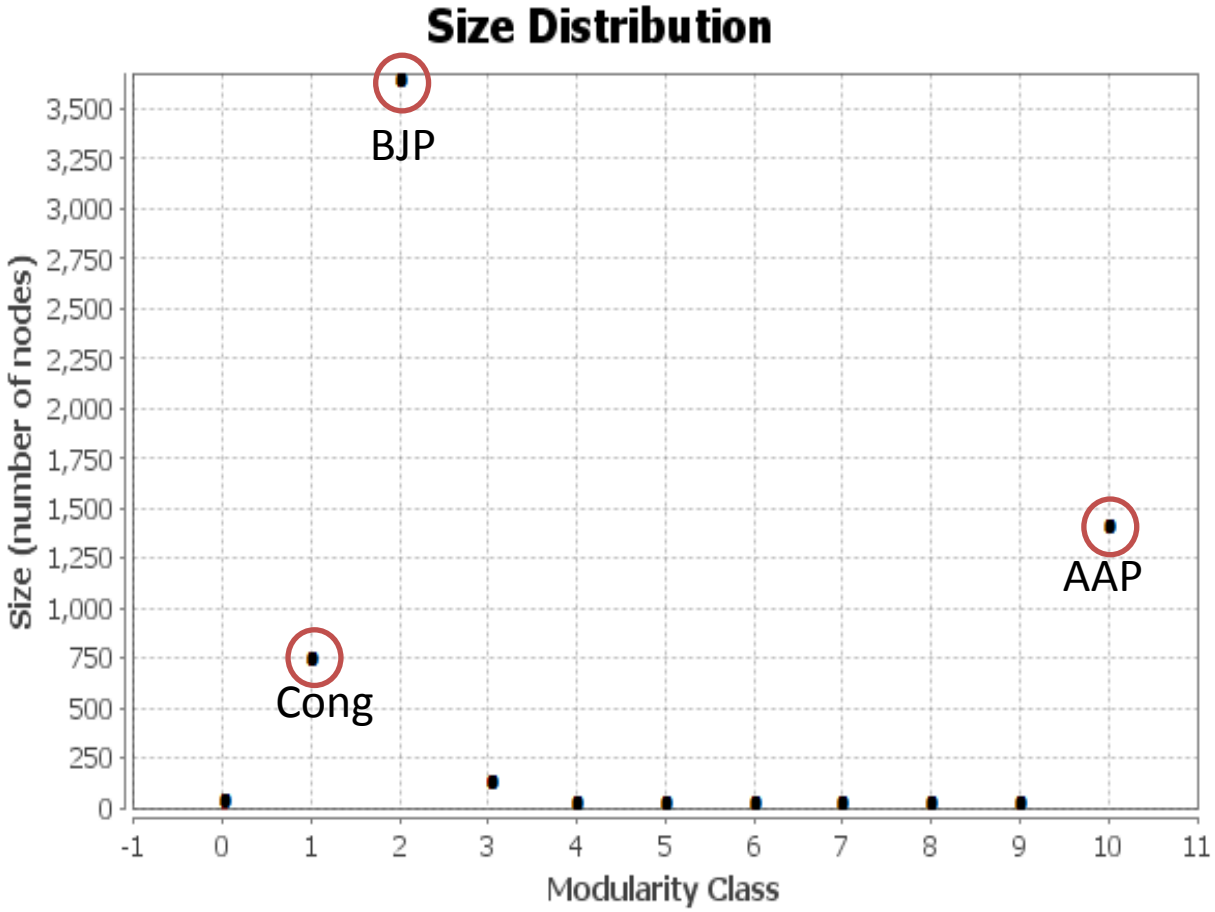
- For BJP-Cong also the efficiency was > 60%
- This method works well for 2-class classification
- For 'Anti' category, there was a 6% improvement with this method

Network Based Classification

- Network formed on the basis of
  - Retweets
  - User mentions
- Undirected, without weights graph
- Users formed the nodes
- Used Gephi 0.8.2
- Community detection algorithm
  - Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks
  - Fastest for large networks

Network Based Classification: Results

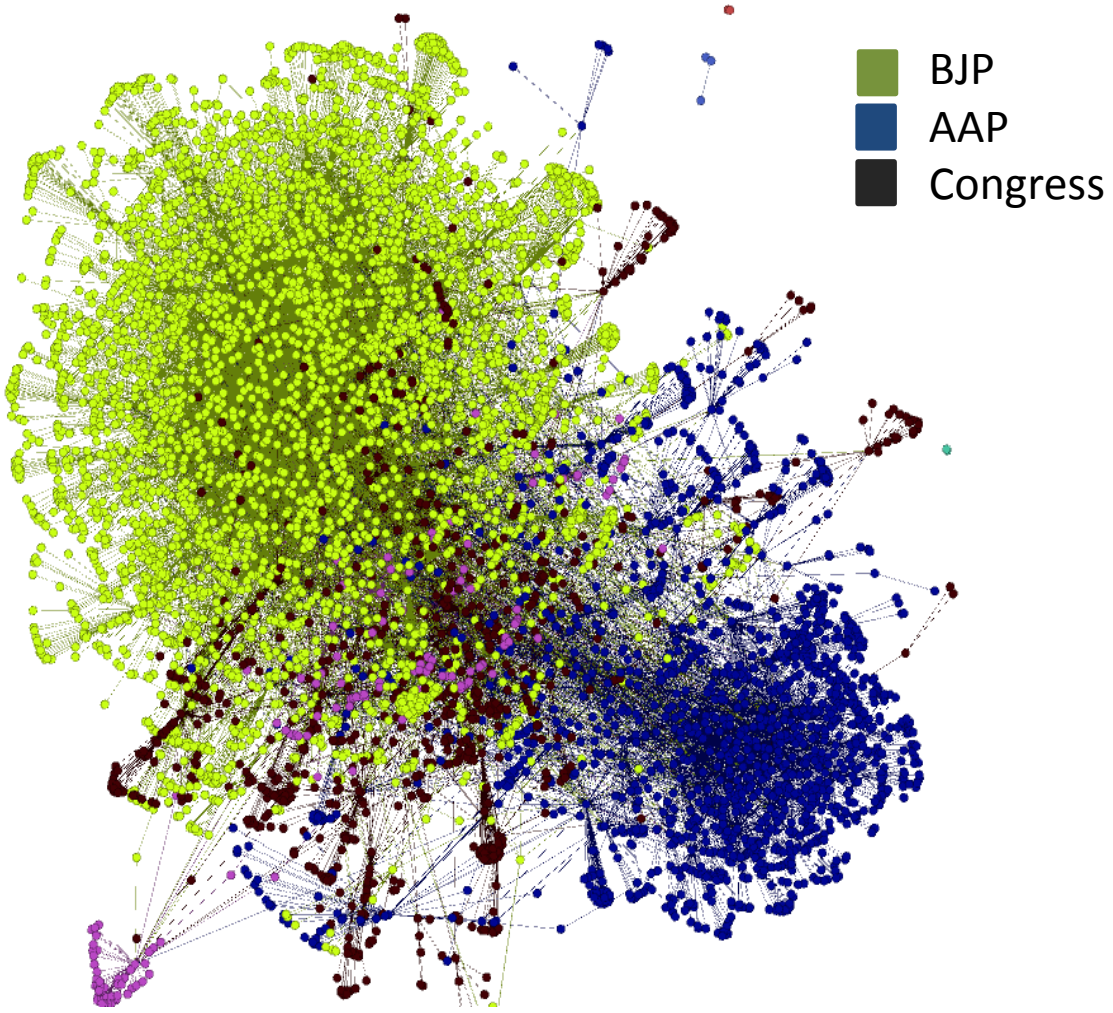
- With all 613 ‘Pro’ instances
- #Communities: 11
- 3 major communities
- Rest had 0.05% of nodes
- Modularity Score: 0.402



Network Based Classification: Results

- With all 613 ‘Pro’ instances

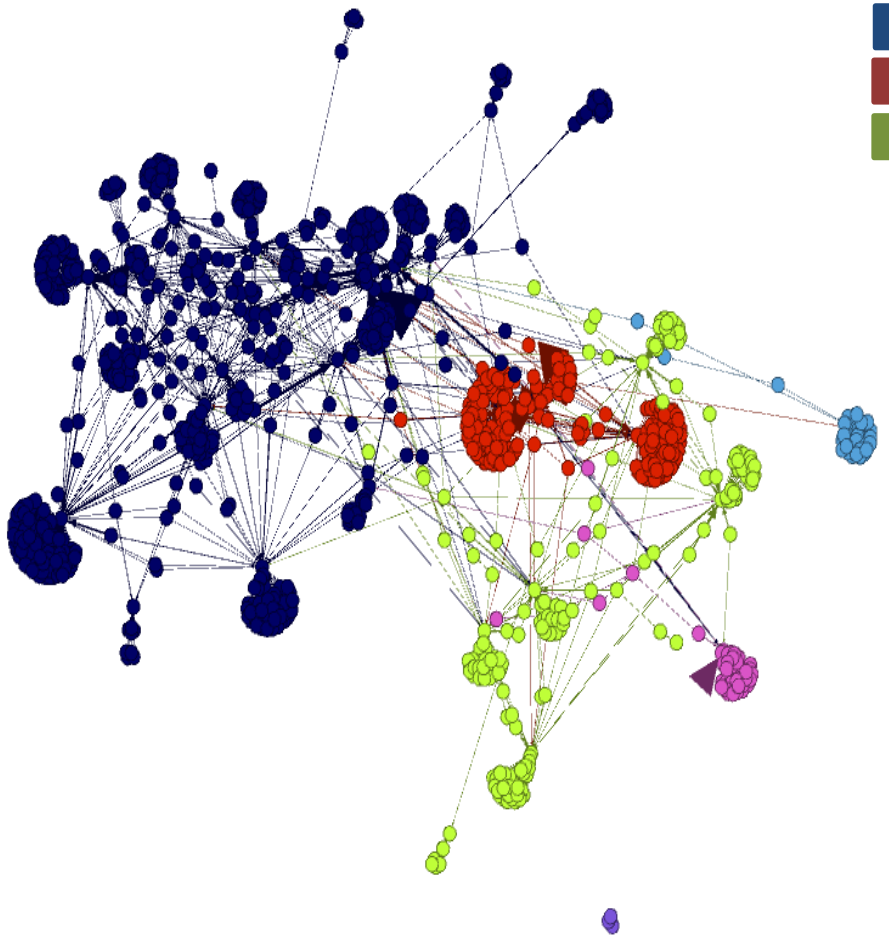
#Nodes: 6022			
#Edges: 13693			
Modularity Score: 0.402			
Efficiency: 78.31%			
Party	**Pre**	**Recall**	**F-measure**
AAP	0.709	0.672	0.690
BJP	0.939	0.850	0.897
CONG	0.326	0.576	0.451



Network Based Classification: Results

- With equal instances of ‘Pro’ category
- #Communities: 8
- 3 major and rest with 0.05% of nodes

#Nodes: 1193			
#Edges: 1489			
Modularity Score: 0.582			
Efficiency: 80.00%			
Party	**Pre**	**Recall**	**F-measure**
AAP	0.856	0.733	0.794
BJP	0.769	0.952	0.860
CONG	0.818	0.897	0.857



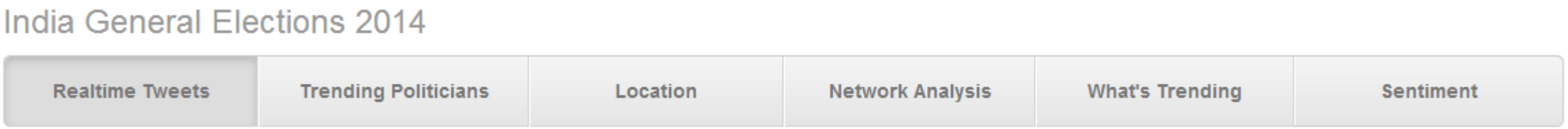
Presentation Outline

- Research Motivation
- Research Aim
- Related Work
- Research Gap
- Data Analysis
- Classification of Political Orientation
- System Design
- Conclusion
- Limitations and Future Work

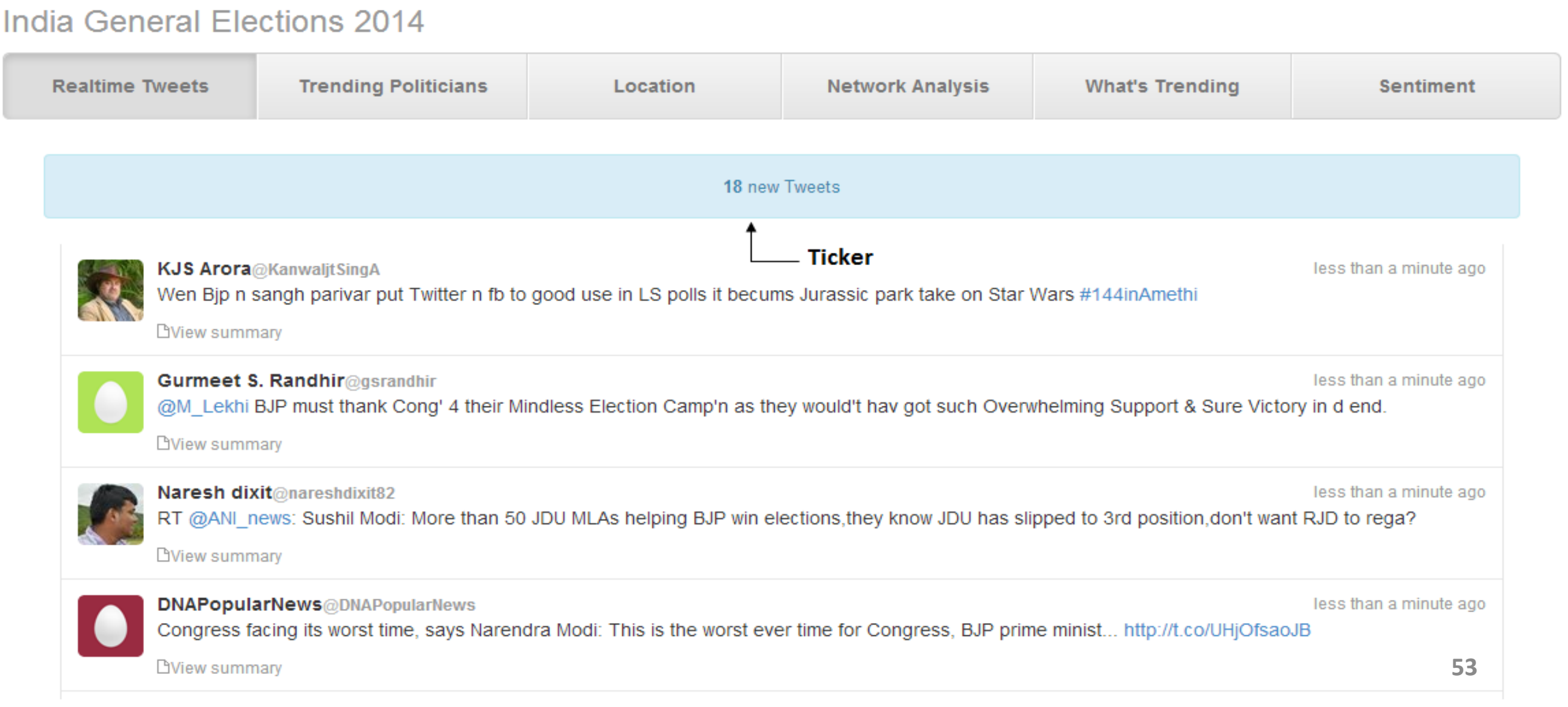
System Design

- Requirement: To be able to see and analyze the election related tweets on daily basis
- Used PHP and Javascript to develop the portal

  http://bheem.iiitd.edu.in/IndiaElections
- Refreshes at a 24 hour interval, but displays the tweets at an interval of 5000 ms



Realtime Tweets



Trending Politicians

India General Elections 2014

Realtime Tweets | Trending Politicians | Location | Network Analysis | What's Trending | Sentiment

Image	Name	Screen Name	Followers	Klout score ▲
	Narendra Modi	narendramodi	3892386	88.9544376106022
	Shashi Tharoor	ShashiTharoor	2158589	83.8670452261145
	Omar Abdullah	abdullah_omar	495903	82.2925779719157
	Rajeev Chandrasekhar	rajeev_mp	127523	80.4229036163932
	Dr Manmohan Singh	PMOIndia	1202471	80.2100714321393
	Milind Deora	milinddeora	131956	79.6566835279628

Location

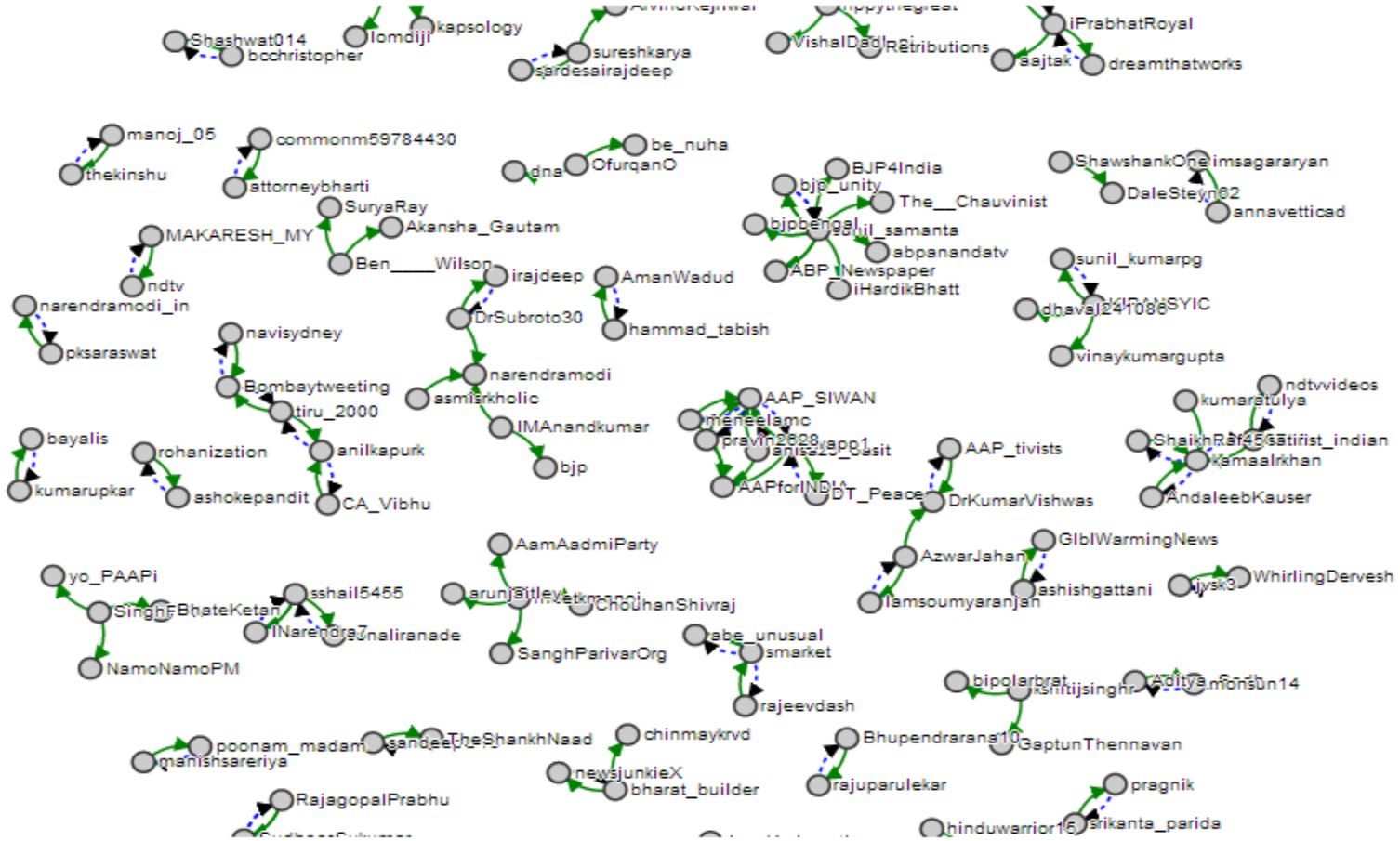
India General Elections 2014

Realtime Tweets	Trending Politicians	Location	Network Analysis	What's Trending	Sentiment

Network Analysis

India General Elections 2014

Realtime Tweets	Trending Politicians	Location	Network Analysis	What's Trending	Sentiment

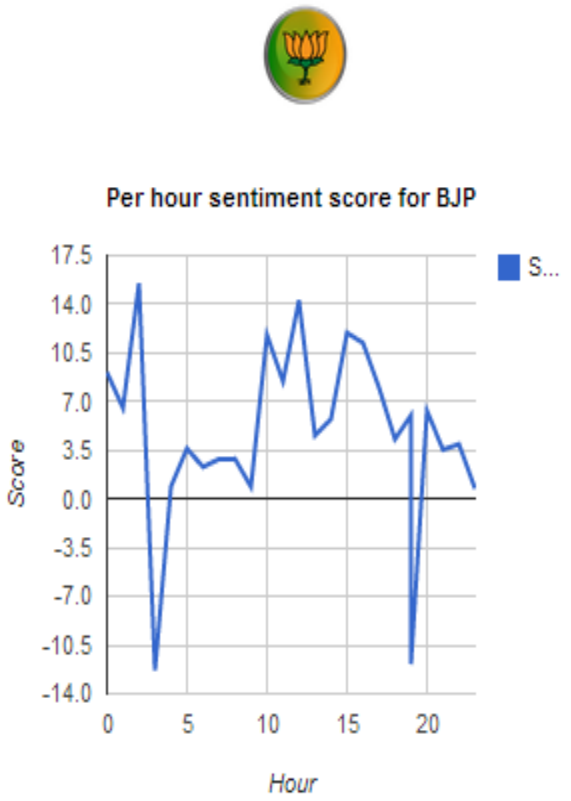


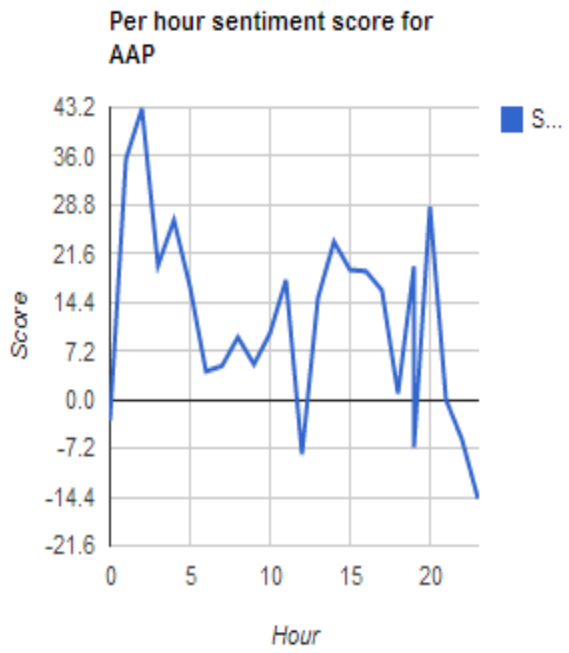
What's Trending

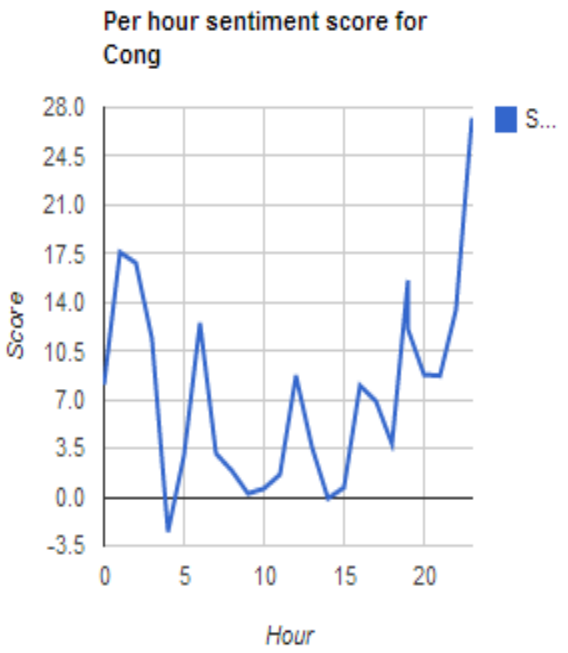
Sentiment

India General Elections 2014

Realtime Tweets	Trending Politicians	Location	Network Analysis	What's Trending	Sentiment







Presentation Outline

- Research Motivation
- Research Aim
- Related Work
- Research Gap
- Data Analysis
- Classification of Political Orientation
- System Design
- Conclusion
- Limitations and Future Work

Conclusion

- Twitter activity is directly proportional to the real time activities
- Data is particularly higher on weekdays
- BJP emerged as the leader in tweet share as well as seat share in both Assembly and General Elections
- Predicting the political orientation with content based methods is not easy
- The transliteration and sarcasm used in the text can be possible causes for poor performance of content based methods

Conclusion

- The user features based classification can improve the efficiency, but not for the ‘Anti’ category
- Prediction of ‘Anti’ political orientation is even more difficult
- Network based methods worked best for the Indian users
- A 2-class classification gives better results in all the methods as compared to 3-class classification
- A system to monitor and analyze the recent tweets was also developed

Presentation Outline

- Research Motivation
- Research Aim
- Related Work
- Research Gap
- Data Analysis
- Classification of Political Orientation
- System Design
- Conclusion
- Limitations and Future Work

Limitations & Future Work

- API rate limit puts a restriction on the data collected
- Political views of people is difficult to judge only on the basis of tweets
- Too much of content by BJP made the results biased towards BJP
- Future work can include to see the change in sentiments post elections
- To look at if the parties that lost the elections were still active and trending

Acknowledgement

- Dr. PK, IIIT-D
- Anupama Aggarwal, PhD Scholar, IIIT-D
- Data Annotators
- CERC@IIIT-D
- Precogs
- Family and Friends

References

[1] Aramaki, E., Maskawa, S., and Morita, M. Twitter catches the u: detecting inuenza epidemics using twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (2011), Association for Computational Linguistics, pp. 1568-1576.

[2] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, 10 (2008), P10008.

[3] Bollen, J., Mao, H., and Pepe, A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In ICWSM (2011)

[4] Boyd, D., Golder, S., and Lotan, G. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on (2010), IEEE, pp. 10.

[5] Bruns, A., and Burgess, J. E. The use of twitter hashtags in the formation of ad hoc publics.

[6] Carletta, J. Assessing agreement on classi_cation tasks: the kappa statistic. Computational linguistics 22, 2 (1996), 249-254.

[7] Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In Proceedings of the 20th international conference on World wide web (2011), ACM, pp. 675-684.

[8] Cohen, R., and Ruths, D. Classifying political orientation on twitter: Its not easy! In Proceedings of the 7th International Conference on Weblogs and Social Media (2013).

[9] Conover, M. D., Gonc_alves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. Predicting the political alignment of twitter users. In Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom) (2011), IEEE, pp. 192-199.

References

[10] Fruchterman, T. M., and Reingold, E. M. Graph drawing by force-directed placement. Software: Practice and experience 21, 11 (1991), 1129-1164.

[11] Gayo-Avello, D. Don't turn social media into another'literary digest'poll. Communications of the ACM 54, 10 (2011), 121-128.

[12] Golbeck, J., and Hansen, D. Computing political preference among twitter followers. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2011),

ACM, pp. 1105-1108.

[13] Gupta, A., and Kumaraguru, P. Credibility ranking of tweets during high impact events. In Proceedings of the 1st Workshop on Privacy and Security in Online Social Media (2012), ACM, p. 2.

[14] Himelboim, I., McCreery, S., and Smith, M. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. Journal of Computer-Mediated Communication 18, 2 (2013), 40-60.

[15] Honey, C., and Herring, S. C. Beyond microblogging: Conversation and collaboration via twitter. In System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on (2009), IEEE, pp. 1-10.

[16] Hong, S., and Nadler, D. Does the early bird move the polls?: the use of the social media tool'twitter'by us politicians and its impact on public opinion. In Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times (2011), ACM, pp. 182-186.

[17] Jungherr, A., J□urgens, P., and Schoen, H. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im predicting elections with twitter: What 140 characters reveal about political sentiment. Social Science Computer Review 30, 2 (2012), 229-234.

[18] Krenn, B., Evert, S., and Zinsmeister, H. Determining intercoder agreement for a collocation identi_cation task. In Proceedings of KONVENS (2004), pp. 89-96.

[19] Metaxas, P. T., Mustafaraj, E., and Gayo-Avello, D. How (not) to predict elections. In privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (SocialCom)

Feedback/ Suggestions:

pk@iiitd.ac.in
abhishek1231@iiitd.ac.in