Can #*Twitter_Trends* Predict Election Results? Evidence from 2014 Indian General Election

Aparup Khatua
Dept. of Comp. Sc. & Eng.
Univ. of Calcutta, India
aparupkhatua@gmail.com

Apalak Khatua
XLRI Xavier School of Management, India
apalak@xlri.ac.in

Kuntal Ghosh
Indian Statistical Institute, Kolkata, India
kuntal@isical.ac.in

Nabendu Chaki
Dept. of Comp. Sc. & Eng.
Univ. of Calcutta, India
nabendu@ieee.org

Abstract

Extant literature finds that twitter trends can capture electoral sentiment. However, empirical evidences are ambiguous in nature. Thus, this study uses the context of 2014 Indian General Election to test the predictive power of Twitter in a large and politically diversified country. We have analyzed roughly 0.4 million tweets during the period March 15, 2014 to May 12, 2014. We observe that tweet volume as well as sentiment analysis can predict election results. We also find that sentiment scores can predict changes in vote share. We note that in a multi-party system the nationality of a party can be an important factor. However, these results should be interpreted with caution. We emphasize the relevance of contextual understanding for efficient data collection and analysis.

1. Introduction

Nowadays social media platforms, like Twitter, are generating enormous amount of user content data. Analysis of this user content data helps to identify small discrete events, like sentiment of an individual user (which might be of less significance), but aggregate level analysis of these discrete individual sentiments can be an efficient indicator of collective trends [2]. Researchers are exploring these voluminous data for studying different socio-economic phenomena.

Explanatory power of tweet feeds for predicting election results is ambiguous in nature [4], [9]. For example, mere volume analysis of political tweets accurately predicted the election result in Germany [15] whereas sentiment analysis failed to predict the 2008 US Presidential election [3], [10]. In spite of these mixed empirical evidences, use of social media data as a predictor of election results is becoming popular among practitioners and researchers. Political parties, even in developing countries, make conscious efforts to manage social media properly during their campaigning phase [1], [11]. Twitter is rapidly gaining popularity in developing countries. There are roughly 33 million registered users in India [11].

This paper uses a new dataset from Indian context to explore the explanatory power of twitter for predicting 2014 Indian general election. We perform both twitter volume as well as sentiment analysis. Broadly our analysis confirms previous findings but it also highlights that a nuanced understanding of research context is extremely important for collecting data. Thus, predicting election results is challenging in a large and politically heterogeneous country like India where both national as well as regional parties (with limited political presence) participate in general election and election is a month long process.

2. Election Prediction & Twitter

Election result prediction is a well-researched area. Studies from disciplines, like sociology or economics, attempt to predict election results through econometric modeling of aggregate level economic or demographic data like GDP growth or unemployment. These studies assume low GDP growth or high unemployment would enhance anti-incumbency effects. However, this approach has limitations in capturing sentiments of individual electorate. Analysis of micro-blogging data like tweet feeds addresses these shortcomings. Twitter allows its users to post/share and read short 140 characters text messages known as tweets [15]. Researchers are analyzing tweet texts for efficient forecasting of election results [3], [4], [10], and [15].

Extant literature explores twitter trends for predicting election results in countries like Dutch [13], USA [3], [9], [10], Singapore [14], Germany [15] and Pakistan [1]. However, these countries are mostly characterized either by a two-party system or a multi-party system with relatively low fragmentation. Indian context is characterized by a multi-party system with relatively high fragmentation (regional parties have strong presence only in certain parts of the country)

1530-1605/15 $31.00 © 2015 IEEE
DOI 10.1109/HICSS.2015.202

[16]. A fragmented party system like India offers a complex context for predicting election results through twitter trends.

In India administrative responsibilities are divided between central government and state governments. In some states the ruling party is same both at the center and the state, whereas for others it might be different. Many a time state governments accuse central government (if it is an opposing party) for skewed resource allocation when the state suffers due to low economic development. There is a significant amount of variations across states in terms of economic development. Some states are industrially more developed than others, whereas some states are still agrarian in nature. Thus, using generic indicators of economic progress in a regression model, to capture anti-incumbency effects, might be problematic for Indian general election. In a bipolar system the opposition party will always enjoy the benefits of anti-incumbency effects. So, US presidential election boils down to a sentiment analysis between Obama and Romney. This might not be the case in India. A regional party can also take hold of anti-incumbency sentiments instead of the opposition party at the national level. In contrast, a national party can capitalize on anti-incumbency sentiments towards regional state government. Conceptually anti-incumbency sentiments towards a state government should get reflected only in state assembly election but in reality it might influence the general election also.

The challenge becomes more intriguing, if we further consider that Indian election is a month long process and political parties take up regional issues in their campaigning considering these election phases. Hence, we argue that this paper not only uses a new dataset from Indian context but also it explores the predictive power of twitter in a relatively complex and diverse political setting with respect to prior studies.

3. Indian General Election 2014

General Elections in India are normally held at an interval of five years for 543 parliamentary constituencies. Election Commission of India, an independent body, conducted the 2014 General Election in nine phases from 7th April to 12th May 2014. This was the longest election in India's history and also one of the largest-ever elections in the world in terms of eligible voters. The cumulative election turnout was 66.4% which is the highest turnout in India till date.

There were two major alliances: BJP led NDA (National Democratic Alliance) and INC led UPA (United Progressive Alliance) in the 2014 General Election. There were few prominent regional parties like AITMC, CPI(M), BJD etc. in Eastern India; AIDMK, DMK (Dravida Munnetra Kazhagam) etc. in Sothern India; BSP, Samajwadi Party, SAD (Shiromani Akali Dal) in Northern India; and MNS (Maharashtra Navnirman Sena), Shiv Sena in Western India (refer Table 1 for details).

Table 1: Leading Parties in Indian Context

	Party	Full Name	Remarks
1	AAP	Aam Aadmi Party	Newly formed party in 2012
2	AIADMK	All India Anna Dravida Munnetra Kazhagam	Regional ruling party in a Southern state
3	AITMC	All India Trinamool Congress	Regional ruling party in an Eastern state
4	BJD	Biju Janata Dal	Regional ruling party in an Eastern state
5	BJP	Bharatiya Janata Party	National opposition party
6	BSP	Bahujan Samaj Party	Regional opposition party in a Northern State
7	INC	Indian National Congress	Leading party of the ruling alliance
8	LEFT	CPI (Marxist) + CPI	Left alliance (CPI: Communist Party of India)
9	SP	Samajwadi Party	Regional ruling party in a Northern State
10	YSRCP	YSR Congress Party	Newly formed party in 2009
11	TRS	Telangana Rashtra Samithi	Regional ruling party in a Southern state

In addition to these major parties, many smaller registered unrecognized parties also participated in this election. 464 political parties contested in 2014 parliamentary election. Interestingly 300-odd parties participated in less than five parliamentary constituencies out of a total of 543 constituencies. However, leading parties like INC or BJP contested from 464 and 428 seats respectively. Another interesting aspect of Indian general election was the presence of independent candidates. There were 3235 independent candidates in 2014 general elections (i.e. roughly 6 independent candidates from each constituency). In addition to this, Election Commission has started one additional option called NOTA (None Of The Above). It allows an electorate to express his disapproval of all candidates. NOTA option has secured 1.1% of total votes which is a significant number in a multi-party system like India. On average there were roughly 15 candidates in each parliamentary

constituency. It is interesting to note that 40 odd candidates contested from the Varanasi constituency. Incidentally this was one of the constituencies from where prime ministerial candidate of NDA contested.

Some of the regional parties have strong presence in their respective regions. For example, parties like AIADMK or BJD contested from 40 and 21 seats respectively. And AIADMK and BJD won 37 (out of 40) and 20 (out of 21) seats respectively. In 2014 general election the third (AIADMK) and fourth (AITMC) largest party, in terms of seat share, were regional parties. There are many parties, like YSRCP, RJD, TRS etc., which might look miniscule from national perspective but they have a very strong regional presence.

A newly formed party called AAP participated in this 2014 election. AAP came into existence in November 2012 as a consequence of social movements. AAP created huge buzz in social media and that also got reflected in our analysis. Immediately before this general election AAP won 28 out of 70 seats in 2013 Delhi Legislative Assembly Election and formed a coalition government. In 2014 general election AAP participated in 432 seats out of 543 total seats. However, they could secure only four seats in one Northern State. We are considering a party as a regional party if their political influence is restricted to limited geographic regions [16]. For brevity, we restrict our discussion mostly to two major alliances (i.e. NDA and UPA). The list of parties which we have considered in this study is summarized in Table 3.

4. Data Collection & Data Cleaning

Data Collection: A prior study collected data by using domain experts to identify core users through manual interventions and identified another set of users who were present in the network of these core users [14]. Is this database a reliable sample to capture the overall electorate sentiment of the country? Or it is a set of people with their own political biases. If this is not a representative sample then it would be "like going to a political rally and sampling the people gathered there, expecting that it will provide an accurate representation of likely voters" [9].

We have collected tweets posted by common people, political candidates, electronic media as well as political parties to overcome the above mentioned bias. We have collected our tweet data by using 'twitteR' [8] and Application Program Interface (API) offered by twitter. We have considered data three weeks prior to phase 1 election to the end of phase 9 election i.e. March 15, 2014 to May 12, 2014. We have identified a set of keywords like names/abbreviation of political parties, key political personalities, important constituency etc. However, we realize that this static set of keywords fails to capture trending topics during this significantly long period of our data collection. Moreover, in a diverse country like India, trending topics are not same all over the country. So, we identified around 15 politically sensitive cities in India. Furthermore, we developed a program to identify top 10 daily trending topics in twitter against particular WOIED (as previously identified). After removing duplicate trending topics (across 15 cities), we add these trending topics (in addition to our core set of keywords) in our search crawler for that particular day. Thus, our set of keywords for crawling tweet feeds was dynamic in nature. This helped us to capture temporal events/sentiments which might not get captured by a static set of keywords. Furthermore, to overcome the rate limit of twitter API we used to run our search algorithm all throughout the day.

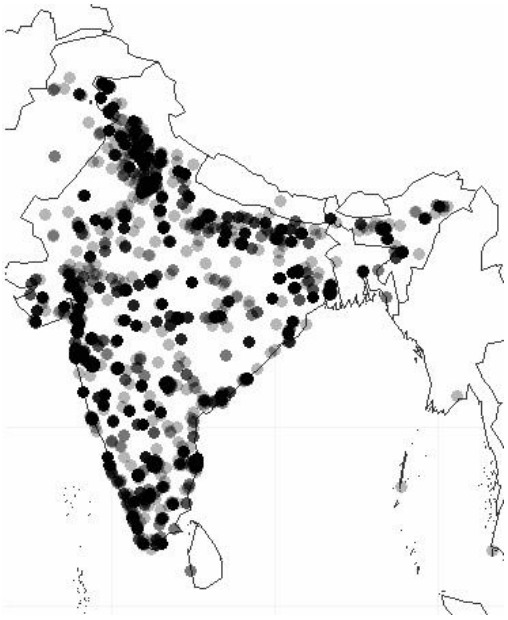

Figure 1: Heat Map of our Tweet Data

Data Cleaning: We have used an exhaustive set of keywords (as described earlier) for crawling purpose. As a consequence some tweets, especially if it contains more than one keyword, got extracted for multiple times. Thus, we apply remove duplicate function on tweet text filed. Moreover, we observe that many tweets are same in terms of text content except a small URL portion. First, we remove these URLs from tweet text then once again we apply the remove duplicate function. Similar to prior study [1], we consider only English tweets for our analysis and discard regional languages tweets. We find that locational data of tweets are available for a small set of tweets. A heat

map of those tweets (refer Figure 1) indicates that our sample has no locational bias and it uniformly represents the Indian electorate. However, the map of Kashmir portion (as generated by *ggplot2* package of R) is debatable.

Table 2: Data Selection Methodology

Sample Tweets	KS 1	KS 2	KS 3	AC	OC
@** ***BJP*** *n* ***@Namo*** *should do well in* ***India***	2	1	-	S-1	RL
@**_** ***TRS*** *Neither Merged With* ***Congress*** *Nor did any alliance with them in* ***Telengana***	2	1	-	S-1	RL
@** ***BJD*** *in* ***Odisha*** *will ensure 20 seats at minimum*	1	1	-	S-1	RL
@*** *Eagerly waiting for* ***NaMo*** *government*	1	0	1	S-2	RL
@**_** *We pay into* ***TRS****(teacher retirement system) but teachers don't get social security*	1	0	0	R-3	J
@**_** *I love the Asian Ball Jointed Dolls, but they are not easily available in UK* ***#BJD***	1	0	0	R-3	J
@** *Having a great time at* ***Varanasi***	0	1	0	R-4	J

KS – Keyword Set; AC – Algorithm Classification; OC – Our Classification; S-2 - Select in Step 2; R-4 – Reject in Step 4; RL-Relevant; J- Junk.

We generate a word frequency list for the entire dataset. Manually we have selected most relevant (KS1) and contextual (KS2) keywords from this word frequency list. We develop two mutually exclusive set of keywords KS1 (like *AAP*, *AIADMK*, *BJD*, *BJP, NaMo, RaGa* etc.) and KS2 (like *Amethi*, *Gujrat*, *India, Varanasi* etc.). KS1 (roughly 430 keywords) mostly comprises of key personalities, party names etc. directly related to 2014 General Election and KS2 (roughly 200 keywords) is broadly contextual in nature like locations, states, politically sensitive constituency etc. Furthermore, we generate another word frequency list from tweets where count of KS1=1 and count of KS2=0. We have manually selected a set of keywords (i.e. KS3, roughly 220 keywords) which have absolutely no ambiguity in terms of relevance (like *abkibaarmodisarkar*, *aapwaveinkashi*, *NaMo, rahulgandhi* etc.). Our algorithm for relevant tweet selection and junk tweets rejection is as follows:

Step 1: If Count (for KS1) >= 1 & Count (for KS2) >=1 then SELECT

Step 2: If Count (for KS1) = 1 & Count (for KS2) =0 & Count (for KS3) =1 then SELECT

Step 3: If Count (for KS1) = 1 & Count (for KS2) =0 & Count (for KS3) = 0 then REJECT

Step 4: If Count (for KS1) = 0 & Count (for KS2) >= 0 then REJECT

Table 2 reports some illustrative examples to elucidate how we have identified relevant (RL) and junk (J) tweets for our analysis. Our manual inspection of a small sub-sample confirms the effectiveness of this algorithm. For example, our algorithm was efficient in identifying junk tweets related to US TRS with respect to actual TRS (refer table 2; R-3). We observe that our algorithm was efficient and stringent in rejecting junk tweets but in the process of removing junk tweets it also rejected relevant tweets. So, there is a trade-off. We feel that selecting relevant tweets is more important for efficient prediction.

5. Hypotheses

Prior studies claim that "the mere number of tweets mentioning a political party can be considered a plausible reflection of the vote share" and can have a high predictive power [15]. Prior studies find that average deviation between tweet share and vote share is insignificant [14], [15]. However, another study finds unacceptably high level of deviation or mean absolute error (MAE) and cautioned about the predictive power of social media data [9]. Broadly prior studies find that "co-concurrence of political party mentions accurately reflected close political positions between political parties and plausible coalitions" [9]. In other words, mere tweet volumes can be an effective indicator of election results. Thus, we propose,

H1A: *Higher tweet share would lead to higher vote share.*

H1B: *Higher tweet share would lead to higher seat share.*

Twitter sentiment analysis is gaining credibility among researchers for capturing public opinions. However, researchers also highlight potential shortcomings of using simplistic bipolar sentiment analysis [5], [9]. A tweet might be in favor of a political party or it might be against a political party. Thus, if we propose that tweet volume analysis is an efficient predictor of vote share then we are implicitly assuming that either all tweets are in favor of the party or the mix of tweets (in terms of favorable and unfavorable tweets) is uniform across parties. These are very strong and not so realistic assumptions. Thus, researchers classify tweets into positive, negative and neutral sentiment. Broadly this stream of research

considers a standard lexicon, used by prior studies [6], [7] for sentiment analysis. A tweet might have a positive or negative score (depending on the number of positive or negative words, as defined by the lexicon, within the tweet). Even a tweet might be classified as neutral if it doesn't has any positive or negative words, or has equal number(s) of positive and negative words (then they cancel each other). This way the individual score of each tweet can be calculated and a summation of the same would give a cumulative sentiment score. Conceptually favorable sentiments regarding a political party would be a better indicator of their electoral performance. Thus, we propose,

H2A: *Higher sentiment score would lead to higher vote share.*

H2B: *Higher sentiment score would lead to higher seat share.*

We carefully observe the trend of prior state assembly or general elections in India. We focus on cases where a significant change occurs in terms of wining seats with respect to previous election. Anecdotal evidences reveal that in India most parties have loyal voters and these voters maintain their loyalty even during not-so-favorable situation. Thus, a minor vote swing (i.e. change in percentage of votes) can lead to a significant change in terms of seat share. For example, in 2009 parliamentary election Congress did significantly well with respect to their performance in 2004 election. Congress won 61 more seats (with respect to 2004 parliamentary election) but their vote share went up by marginal 2% [12]. In the same parliamentary election Communists faced one of their worst defeats. They won just 24 seats in 2009 election (with respect to 59 seats in 2004 election). Interestingly AITMC, a relatively newer party then, had displaced Communists in West Bengal which was a Communist bastion for many years. AITMC won 17 seats more than 2004 election though the vote share went up by marginal 1.4% [12]. Two years later Communist (which was the ruling state government) had a landslide defeat during 2011 State Assembly Election to AITMC. Communists managed to retain just 60 seats (whereas they had 227 seats in the previous state assembly election) out of 294 seats. Interestingly the drop in their vote share (48% in 2006 to 40% in 2011) was not so drastic. These anecdotal evidences indicate that mostly voters are loyal in Indian context. We argue that minor vote swing by not-so-happy electorate can play a crucial role for replacing the incumbent government. Sentiment scores can effectively capture the mood of this not-so-happy electorate. Thus, we propose,

H3: *Sentiment score would be proportional to change in vote share.*

6. Data Analysis and Findings

Prior studies find that "tweets containing the names of both candidates" can be misleading for forecasting analysis. So, they focus "only on tweets mentioning one candidate at a time" [9]. A tweet which has joint mention of two candidates/parties can be problematic especially for sentiment analysis. For example, a tweet as follows:

#ABC said 'Yes We Can'. Now XYZ would say to his party workers 'No We Can't'☺

Simplistic sentiment analysis would fail to categorize it properly and might consider it as a neutral statement (due to presence of both *Can* and *Can't* in the same sentence) for both ABC and XYZ. However, this tweet is positive for ABC and sarcastically negative for XYZ. Thus, we discard tweets which have joint mention of two parties. For example, if a tweet has any NDA related keyword like *NaMo* as well as mention of any other keywords (related to other parties) from the list of keywords provided in Table 3 then we consider it as a problematic tweet. Thus, we define a tweet for NDA as: tweets which will have only NDA related keywords and there should not be any keywords (which are related to other political parties) from Table 3.

Table 3: Parties/Alliances & Analysis Keywords

		List of Keywords
1	NDA*	NDA,BJP,LJP,advani,modi,#modi, narendra, #narendra, namo, ShivSena, TDP, navnirman, Thackeray, MNS, paswan, Shiromani, & Akali.
2	UPA**	UPA, NCP, Gandhi, Rahul, pappu, Sonia, Robert, Vadra, manmohan, Cong, RJD, lalu, JMM, IUML, INC, raga, & congress.
3	AAP	AAP, #AAP, AamAadmi, #AamAadmi, Arvind, Kejriwal, AAAP, & yogendra.
4	AIADMK	AIADMK, Jayalalithaa, & amma.
5	AITMC	TMC, didi, Mamata, AITC, & trinamool.
6	BJD	BJD,Naveen, & Patnaik.
7	BSP	BSP,Bahujan,& Mayawati.
8	LEFT	CPI(M), CPI, CPM, & Buddhadeb.
9	SP	Samajwadi, mulayam, & akhilesh.
10	YSRCP	YSRCP, Jagan & YSR.
11	TRS	TRS, Chandrasekhar, & telengana.
12	Others	JD(U), Nitish, Anna Hazare, Yeddyurappa, & karunanidhi.

** This was alliance between BJP (main opposition party), SS(ShivSena), TDP (Telugu Desam Party), LJP (Lok Janshakti Party) and SAD (Shiromani Akali Dal) and few other smaller parties.*

*** This was the alliance between INC (the ruling party), NCP (Nationalist Congress Party), RJD (Rashtriya Janata Dal), JMM (Jharkhand Mukti Morcha) and IUML (Indian Union Muslim League) and few others.*

We observe that removing these kinds of tweets from our original dataset, following prior studies [9], [13], has improved the predictive power of our analysis. Following extant literature [1], [14], [15] we select a set of keywords like prime ministerial

candidate's name, prominent leaders from ruling and opposition parties, key political personalities. Details of these keywords are listed in Table 3.

Table 4: Percentages of Votes, Seats and Tweets

	Vote (%)	Seat (%)	Raw Tweet (%)	Tweet (%)	Error 1	Error 2
	(1)	*(2)*	*(3)*	*(4)*	*(4-1)*	*(4-2)*
NDA	0.36	0.60	0.51	0.47	0.11	-0.13
UPA	0.22	0.11	0.32	0.25	0.03	0.14
BSP	0.04	0.00	0.02	0.02	-0.03	0.02
LEFT	0.04	0.02	0.01	0.00	-0.04	-0.01
AITMC	0.04	0.06	0.02	0.01	-0.03	-0.05
SP	0.03	0.01	0.01	0.01	-0.02	0.00
AIADMK	0.03	0.07	0.01	0.01	-0.02	-0.06
YSRCP	0.03	0.02	0.01	0.01	-0.02	-0.01
AAP	0.02	0.01	0.22	0.20	0.18	0.20
BJD	0.02	0.04	0.01	0.00	-0.01	-0.03
TRS	0.01	0.02	0.01	0.00	-0.01	-0.02
Total	0.85	0.95	1.15	1.00		
MAE					4.50%	5.99%

Source for Vote (%) and Seat (%): http://eci.nic.in/eci/eci.html

Table 4 reports party-wise overall outcome of the 2014 general election, in terms of vote share (% of total votes), seat share (% of total wining seats out of 543 seats) and tweet volumes (during the period March 15, 2014 to May 12, 2014). We have complied Table 4 from Election Commission website as well as from our primary tweet data. Figures are approximated to two decimal places. Our final sample of 11 parties/alliances (sorted in terms of vote share) considers vote share, seat share as well as their shares in overall tweet feeds. Our final sample accounts for roughly 85% of total votes and 95% of parliamentary seats.

We observe that AAP (22%) had created enough buzz in social media. Similar to PTI (Pakistan Tehreek-i-Insaf) of Pakistan [1], prominent leaders of AAP had significant number of followers in social sites like Twitter. APP had hired digital marketing agency to evaluate their social media performance [11] and ensured greater interaction with the public. However, this did not get reflected in their overall election performance in terms of vote share (2%) or seat share (1%). So, AAP is a significant outlier in terms of tweet share and vote/seat share relationship in our study.

In contrary, there is a party like BSP which has decent vote share (4%) but that did not get reflected in tweet feeds (2%) as well as seat share (interestingly not even a single seat won by them). Thus, to have a holistic representation of Indian political parties/alliances we consider all three relevant parameters: vote share, seat share and tweet share. In a multi-party scenario number of seats won has more relevance in forming the government than vote share. For example, a regional party (like BSP) from a larger state might have decent vote share (4 %) but that might not get reflected in number of seats won by them. However, a regional party (like BJD) from a relatively smaller state might not have significant vote share (2%) in national context but they did fairly well in terms of seats won (20 seats i.e. 3.7% of all seats) in their respective region.

It is worth to note that the cumulative Raw Tweet percentage is 115% (Table 4, Column 3). We have collected roughly 0.6 million of raw tweets from 0.13 million of unique users. As we discussed earlier, many tweets mention two competing political parties. For example, a tweet can be: 'Party A is better than Party B'. So, in Raw Tweet calculation this is getting counted for both Party A and Party B. That is why the cumulative figure of tweet share is more than 100% in Column 3. Column 4 (of Table 3) reports the tweet share of these leading parties after removing tweets (roughly 0.2 million tweets got discarded here) which have joint mention of two parties. Two leading alliances NDA and UPA cumulatively account for 58%, 71% and 72% of vote, seat and tweet share respectively in our sample. This indicates a duopoly like situation in the competitive landscape of Indian politics.

A simplistic mean average of error reveals that tweet volume analysis has an error of 1.37% for vote share and 0.42% for seat share. However, a careful analysis of few parties like NDA, UPA and AAP shows significant aberrations. Thus, we refrain from making a conclusive statement. It is interesting to note that Error 2 (Tweet Share - Seat Share) for NDA and UPA are respectively -0.13 and 0.14 respectively. Because of their opposite signs they cancel each other and bring the error to a respectable 0.42%. This average might not be the correct indicator for testing the predictive power of twitter. Refer [4] for a detailed discussion. Thus, considering an absolute of error for calculating MAE (Mean Absolute Error) would be a better indicator if we want to test the predictive power of tweet volume analysis [14], [15].

$$MAE = \sum |Error\ of\ NDA| + |Error\ of\ UPA| + \ldots$$

This would not allow errors like -0.13 and 0.14 to nullify each other. We observe a MAE of 4.50% for vote share and 5.99% for seat share. These are unacceptably high. Thus, we argue that it is important to have a cautious approach before considering MAE statistics for justifying predictive power of twitter for forecasting election results [4].

Next we employ Ordinary Least Square (OLS) regression technique following extant work. However, our sample size (11 data points) is a serious limitation for OLS analysis. It is difficult to have a bigger sample in a single country study. Thus, prior studies [14] also use smaller sample size for national-level analysis. We consider both tweet volume as well as sentiment score in our OLS analysis for predicting vote and seat share.

We consider two *Dependent Variables* for H1 and H2 as follows: *Vote Share* i.e. percentage of votes received by a political party/alliance (for H1A and H2A) and *Seat Share* i.e. percentage of seats won by a political party/alliance (for H1B and H2B). For hypothesis 3 we consider the *Changes in Vote Share* as the difference between 2009 General Election and 2014 General Election. However, some of the alliance members of NDA and UPA were different in 2009 General Election. We have considered the same in our calculation. We have extracted data from ECI website for these variables.

We have two *Explanatory Variables* as follows: *Tweet (%)* and Sentiment. We consider tweets which have mention of only one party. We have roughly 0.4 million tweets like this. Thus, Tweet (%) is calculated as party related tweets divided by total 0.4 million tweets. For example, if Party A receives 0.1 million tweets, *Tweet (%)* for party A would be 0.25 (refer Table 4).

For sentiment analysis we have combined two lexicons: Hu & Liu's opinion lexicon [7] and list of AFINN-111 [6]. Both these lists comprise of positive and negative words. Lexicon [7] just classifies words into positive and negative categories whereas the [6] lexicon classifies words from strongly negative (a score of -5) to strongly positive (a score of +5) categories. We add all words, in the range +1 to +5 (-1 to -5), from the lexicon [6] as positive (negative) words in Hu & Liu's opinion lexicon [7]. In other words, we enhanced Hu & Liu's opinion lexicon [7] by considering very positive (negative) of [6] as only positive (negative) words and added the same to [7]. We find that combining these two lexicons has improved the predictability power of our analysis.

However, we would like to caution that considering bipolar sentiment score for predicting election results might be problematic in our context. In a two-party system bipolar score can be a good proxy for electoral sentiments but in a multi-party system with high fragmentation it might not be same. Volume of tweets would be higher for a national party (contesting in all seats) with respect to a regional party (contesting in limited number of seats). For example, 100 negative tweets for a Party A, which has regional presence, would lead to a sentiment score of -100. Here all tweet feeds are negative which indicates none of them are supporting Party A. If we assume that these tweet feeds as a representative sample then the hypothetical vote share for Party A would be 0.0%. If another Party B, which has national presence, gets 900 positive tweets as well as 1000 negative tweets. The sentiment score for Party B would be also -100 but in reality 47% of tweets are in favor of Party B.

Parties which accounts for higher tweet share would normally enjoy higher sentiment score. For example, Party A, which has regional presence, getting 20 positive tweets and 10 negative tweets. Thus, cumulative sentiment score would be 10. If another Party B, which has national presence, might get 200 positive tweets and 100 negative tweets. Here sentiment score for Party B is 100, which is higher than the score of Party A. However, 2/3rd tweets were positive for both the parties. Hence, sentiment scores in a fragmented political system should be interpreted with caution. Moreover, *Tweet (%)* is a normalized data whereas sentiment scores are not normalized in nature. To tackle this problem first we calculate the bipolar sentiment score (*Sent_Score*) using our combined lexicon database. Next, we compute our *Sentiment* variable as individual sentiment score of each party/alliance divided by the cumulative sentiment score of all 11 parties/alliances. The standard deviation came down to an acceptable range for *Sentiment.*

Control Variable: Our study incorporates both national as well as regional parties in terms of their political presence. Thus, to capture this aspect we develop a variable called *Nationality*. We computed *Nationality* of a party by dividing the number of seats in which a party contested by 543 (i.e. the total number of parliamentary seats). Thus, a national party/alliance like UPA which is contesting all over India would have a score of 1.0 whereas a regional party like AIDMK (which contested only in 40 seats) would have a score of 0.07. However, we didn't find any significant results for our *Nationality* variable in our OLS regression analysis (for brevity we have not reported these results in the paper). We carefully explore this issue and observe that parties like BSP or AAP contested from 503 and 432 seats respectively but these parties have limited political presence. For example, AAP won only four seats in one north Indian state whereas BSP failed to secure a single seat in this election. Thus, the variable *Nationality* lost its relevance because of few aberrations like this. So we computed a new variable called Nationality Dummy (*Nat_Dum*). Here we assign a value of 1 for parties/alliances like NDA and UPA which have national presence and assigned a value zero for all other parties with limited geographical presence. This conceptualization is similar to prior studies in Indian context [16].

Table 5: OLS Regression for predicting Vote Share and Seat Share

	DV: Vote Share (M1 to M6)						***DV: Seat Share*** (M7 to M12)					
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
Tweet (%)		**0.66*****	**0.24***			**1.12*****		**0.96*****	**0.85***			**1.07****
		(0.10)	(0.13)			(0.14)		(0.19)	(0.42)			(0.45)
Sentiment				**0.40****	0.12	**-0.42*****				**0.68*****	**0.41***	-0.11
				(0.15)	(0.06)	(0.12)				(0.21)	(0.21)	(0.37)
Nat_Dum	**0.27*****		**0.18*****		**0.24*****		**0.33*****		0.05		**0.22****	
	(0.03)		(0.05)		(0.03)		(0.09)		(0.16)		(0.10)	
Constant	0.03**	0.02	0.02*	0.04	0.02*	0.01	0.03	-0.00	0.00	0.02	0.01	-0.00
	(0.01)	(0.02)	(0.01)	(0.03)	(0.01)	(0.01)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)
R^2	0.91	0.83	0.94	0.44	0.94	0.94	0.59	0.73	0.73	0.54	0.73	0.73
Adj. R^2	0.90	0.82	0.93	0.38	0.92	0.92	0.55	0.70	0.67	0.49	0.66	0.66
F-Stat	94.2***	45.4***	63.1***	7.2**	61.0***	60.3***	13.1***	24.2***	10.9***	10.7***	10.7***	10.9***
VIF	1.0	1.0	4.13	1.0	1.45	4.83	1.0	1.0	4.13	1.0	1.45	4.83

N=11; Two-tailed tests; Standard error in parenthesis ;< 0.10 ** p<0.05, *** p <0.01; VIF - Variance Inflation Factor*

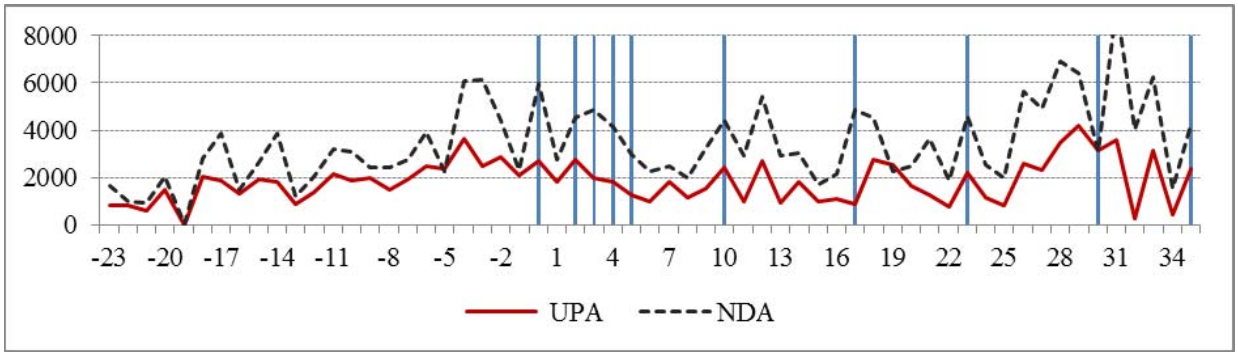

Figure 2: Tweet Volumes of UPA and NDA during March 15, 2014 to May 12, 2014

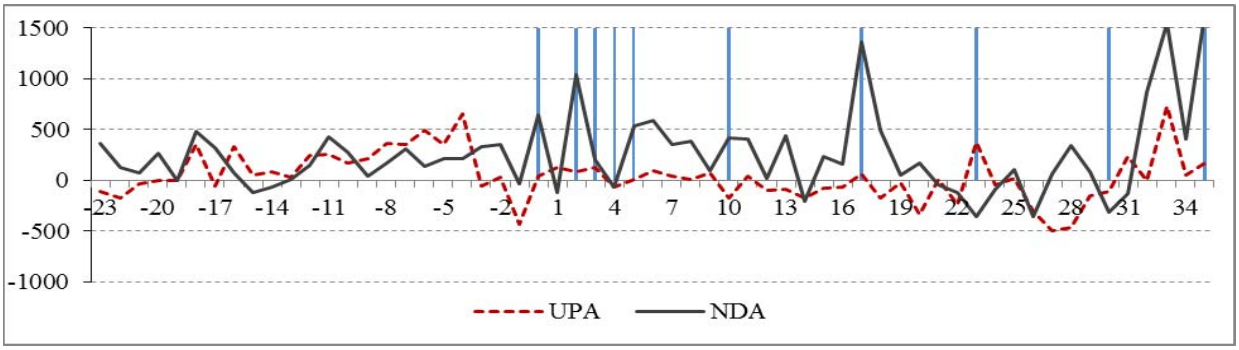

Figure 3: Sentiment Score of UPA and NDA during March 15, 2014 to May 12, 2014

Table 5 reports the OLS regression results for H1and H2. We find that coefficients of our *Nat_Dum* variable are positive and statistically significant for almost all models (in Table 5) i.e. a national party would have a higher propensity to win more seats or secure more votes than a regional party which is intuitive. We observe that the coefficients of *Tweet (%)* are positive and statistically significant for all models. The results remain consistent even after incorporating the nationality/regional factor (i.e. *Nat_Dum*) in our analysis. Thus, it *strongly supports Hypotheses H1A and H1B*. So, higher tweet share leads to higher vote share as well as higher seat share. Similarly coefficients of *Sentiment* are positive and statistically

significant for most models. This indicates higher sentiment score leads to higher vote and higher seat share. But interestingly the coefficient of Sentiment for predicting vote share becomes insignificant when we are incorporating the control variable *Nat_Dum* (Table 5, Model 5). Thus, our OLS analysis *weakly supports our Hypotheses H2A and H2B*.

However, findings are ambiguous when we are incorporating both *Tweet (%)* and *Sentiment* in our forecasting model (Table 5, Models 6 and 12). Results for *Tweet (%)* remain consistent but coefficient of *Sentiment* becomes negative and significant for vote share analysis (Model 6 of Table 5) and statistically insignificant for seat share analysis (Model 12 of Table 5). Problems related to sentiment score in a fragmented political system, as we discussed earlier, might be a reason. However, this requires further exploration. We find significant multicollinearity issues when we incorporate *Tweet (%)*, *Sentiment* and *Nat_Dum* in the same regression model (for brevity we have not reported it).

Table 6: OLS Regression for Vote Swing

	M1	M2	M3	M4
Sent_Score	**0.00****		**0.00*****	
	(0.00)		(0.00)	
Sentiment		**0.22****		**0.48*****
		(0.09)		(0.07)
Nat_Dummy			**-0.13*****	**-0.13*****
			(0.03)	(0.03)
Constant	-0.01	-0.01	-0.00	-0.00
	(0.02)	(0.02)	(0.01)	(0.01)
R^2	0.46	0.46	0.89	0.89
Adj. R^2	0.39	0.39	0.85	0.85
F-Stat	6.0**	6.0**	24.4***	24.3***
VIF	1.0	1.0	2.4	2.4

N=9; Two-tailed tests; Standard error in parenthesis ;< 0.10*
***p<0.05, *** p <0.01; VIF - Variance Inflation Factor*

Figure 2 and figure 3 plot tweets volumes and sentiment scores respectively for two leading alliances NDA and UPA. Blue vertical lines, both in Figure 2 and Figure 3 indicate polling dates across the country. Day 0 (April 7, 2014) was the polling date for first phase election whereas Day 35 (May 12, 2014) was the last polling date i.e. ninth phase election. Graphical plots, both tweet volumes and sentiment scores, are quite indicative of the electoral performance of NDA and UPA. It would be interesting to probe a distinct negative spike of UPA graph during the eighth (around day 26/27) election phase. We observe that during eighth phase (i.e. day 30 in Figure 3) election took place in 64 seats. In this phase UPA won in only one seat (in Amethi, Uttar Pradesh) whereas NDA won in 49 seats out of 64. This confirms the negative spike in UPA sentiment score around day 26/27 in Figure 3.

Next we explore the relationship between sentiment score and vote swing. We use two measures: non-normalized *Sent_Score* and normalized *Sentiment*. Coefficients for both variables are positive and statistically significant (Models 1 and 2 of Table 6). This result remains consistent even after controlling the nationality of a party. Coefficients of *Nat_Dummy* are negative and statistically significant (Models 3 and 4 of Table 6). This indicates that chances of vote swings are lower for national parties (i.e. their vote share is stable) and higher for a regional party in Indian context. Overall our findings *strongly support Hypothesis 3*. The sample size, for this analysis, is smaller because parties like YSRCP and AAP were not formed during 2009 general election.

8. Conclusions

This study has made an attempt to answer whether twitter trends can predict election results. Our research setting in Indian context offers a complex political landscape. For example, in India vote share might not be a good proxy for number of seats own. We have highlighted contrasting cases of BSP and BJD. Thus, it is important to test the explanatory power of twitter trends in predicting vote as well as seat share separately like [13]. Our findings broadly confirm prior studies. However, it also illuminates few potential pitfalls of using simplistic kind of analysis like average errors versus absolute errors. Our contributions are in number of fronts.

First, our study has made an attempt to develop a kind of template for data collection and cleaning which can be used by researchers for similar kind of work. Our domain driven data mining model has significantly improved our data collection and relevant tweet identification process. We argue that nuanced contextual understanding is essential for rejecting junk tweets and selecting relevant tweet feeds. Future research should explore how to improve the trade-off between rejecting junk and selecting relevant tweets.

Second, in countries like India where regional parties have strong presence in their respective constituency, it is important to consider and control regional dynamics for efficient prediction. We have made an attempt in this direction (by incorporating a control variable *Nat_Dum*) but this can be a potential research area for future studies.

Third, our findings regarding tweet volume and sentiment score confirm prior studies. However, it leads to ambiguous outcome when we are using both of them simultaneously in the same model. This requires

further exploration in other contexts. Probably our study is fraught with the pitfall of small sample size.

Fourth, we find that significant portion of voters are loyal to a political ideology, sentiment score can be an effective predictor of vote swing. Our empirical evidences support this proposition.

A better sentiment analysis might help us to do a fine grained analysis between vote share and changes in vote share, which is a limitation of this study. We demonstrate with few hypothetical examples why sentiment scores might be misleading in predicting elections results in countries like India. In a large country like India where election is a month long process it might make sense to consider a shorter timespan for data collection. Probably a phase-wise approach of data collection might be a better predictor of election results for regional parties. Broadly our paper confirms the explanatory power of twitter trends for predicting election results. However, future studies should juxtapose twitter trend analysis with professional pollsters to conclusively answer whether twitter trends can predict election results more efficiently than other methods [5], [9], [15].

Acknowledgements: This publication is an outcome of the R&D work undertaken in the ITRA project of Media Lab Asia in the areas of Mobile Computing, Networking and Applications (ITRA-Mobile).

8. References

[1] Ahmed, Saifuddin, and Marko M. Skoric. (2014). My name is Khan: the use of Twitter in the campaign for 2013 Pakistan General Election. *System Sciences (HICSS), 2014 47th Hawaii International Conference on. IEEE*, 2014.

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

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

[4] Gayo-Avello, D. (2012a). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *arXiv preprint arXiv:1206.5851*.

[5] Gayo-Avello, D. (2012b). "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"--A Balanced Survey on Election Prediction using Twitter Data. *arXiv preprint arXiv:1204.6441*.

[6] Hansen, L. K., Arvidsson, A., Nielsen, F. Å., Colleoni, E., & Etter, M. (2011). Good friends, bad news-affect and virality in twitter. In *Future information technology* (pp. 34-43). Springer Berlin Heidelberg.

[7] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.

[8] Gentry, J. (2013). twitterR: R based Twitter client. R package version 1.17, URL http://CRAN.R-project.org/package=twitterR

[9] Metaxas, P. T., Mustafaraj, E., &Gayo-Avello, D. (2011). How (not) to predict elections. In Privacy, security, risk and trust (PASSAT), *2011 IEEE third international conference on social computing* (SocialCom) (pp. 165-171). IEEE.

[10] O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *ICWSM*, 11, 122-129.

[11] Patel, A. (2014) India's social media election battle, URL http://www.bbc.com/news/world-asia-india-26762391,

[12] Sadanandan, A. (2009). The parliamentary election in India, April–May 2009.*Electoral Studies*, *28*(4), 658-662.

[13] Sang, E. T. K., & Bos, J. (2012). Predicting the 2011 Dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media* (pp. 53-60).Association for Computational Linguistics.

[14] Skoric, M., Poor, N., Achananuparp, P., Lim, E. P., & Jiang, J. (2012). Tweets and votes: A study of the 2011 Singapore general election. In *System Science (HICSS), 2012 45th Hawaii International Conference* on (pp. 2583-2591).IEEE.

[15] Tumasjan, A., Sprenger, T. O., Sandner, P. G., &Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM*, 10, 178-185.

[16] Yadav, Y. (1999). Electoral Politics in the Time of Change: India's Third Electoral System, 1989-99. *Economic and Political Weekly*, 2393-2399.