Forecasting Political Weather for Pakistan Local Government Election-Using Opinion Mining

Rahman Ullah
MS Research Scholar
ICIT, Gomal University,
Dera Ismail Khan, Pakistan
rahman_lakki@yahoo.com

Abdur Rashid Khan
Professor ICIT, Gomal University,
Dera Ismail Khan, Pakistan
rashidkh08@yahoo.com dr.arashid@gu.edu.pk

Muhammad Irfan
MS research Scholar, ICIT,
Gomal University, Dera Ismail Khan
irfan@yahoo.com

Abstract- Microblogging websites have changed the methods of information sharing with other users and terminals of the network. The onlookers of information not merely utilize the information at hand but manipulate and process to generate the information of their own interest. Social media especially twitter is becoming famous for political weather forecasting. During the election time many researchers trap the real time data and analyze to generate information of their own interest. In this paper we investigate that how we could use twitter data for forecasting Pakistan local government election in the federal capital Islamabad. The accuracy of the proposed model was 0.330. Statistics showed that only 10% of Islamabad residents were registered on twitter. So it's not feasible to predict the whole population by this small figure. The gap between predicted and actual results can be minimized for large and more refined samples.

Keywords— Sentiment Analysis, Elections Predictions, Opinion Mining, Blog Mining

I. INTRODUCTION

Weblogs are best dialogic forum for collecting information from multiple dimensions regarding political discussions as compared with traditional media and [1, 2]. It is empirically investigated that discussion boards are the best sources for political discussion evaluation with added capability of highlighting important information and user interaction with blog based discussions [3, 4, 5].

Political blogs have great impact on world politics by supervising and organizing the followers but this digital content is not sufficient to cover the offline discussions about politics studied that social media users like Facebook can predict the election results [6, 7]. A paradigm of political opinion can be sketched from the tweets shared in weblogs, but still no analytical mechanism exists to trace the political feelings in microblogs [8]. Though more simple calculations yields more amazing results such as [9] publicize that press members predict better than the election polls observed the association pattern between the blog members reveal their opinions along the party policies [10].

In election days people performed research in multiple areas showed that twitter is used in US 2010 election [11]. The area of interest for incumbents is ongoing events and challengers used twitter for the criticism of incumbents. Researchers highlighted that these tweets called the users to do something and inspired the users to utilize their vote and so twitter also play important role in the voter mobilization [12].

These researchers showed that in political discourse twitter play a vital role and large data chunks are found to analyze the political opinion and behavior of the voters. Hashtags are used to mention the hot topics the users are actively involved in, and other users will mark these topics if they noticed [13]. Twitters completely confine the political happenings of the world and mention people's mood [14].

II. LITERATURE REVIEW

The social media means are used to collect the people's opinion about any topic going on. Talking about anything highlights the number of domains where opinion mining can be applied; it marks the differences between different domains for example movie reviews, drug reviews, airline reviews, hotel reviews, product reviews and political reviews. So it becomes clear that different kinds of tasks and techniques used in opinion mining may vary for different domains.

In recent years opinion mining is also active in political concerns. Political parties and their candidates need to know and read the people's mind and to compute their popularity during the election times. Opinion mining is important from government perspective in terms to track the candidate's status. Plenty of research has been conducted in area of sentiment analysis and opinion mining to extract user thoughts shared in the Weblogs.

Supervised learning techniques were applied to analyze political blogs and categories sentiment about any topic in the political weblogs [15]. The posts can be classified in to positive and negative classes and Naïve Bayesian (NB) classifier results are better than Support Vector Machine (SVM). They noticed that reducing 30 % average dataset size for the purpose to balance classes doesn't affect the total accuracy negatively. The misclassified posts can be categorized in a better manner". They analyzed the key challenges crossing the way while studying political opinion mining such as the variable writing styles used and different speech forms, so it became difficult to analyze the opinion under Sarcasm and cynicism.

The lexical knowledge and the techniques of text classification were applied to extract opinion in political blogs discussing candidates of the US presidential election, products reviews and online movie reviews [16]. They suggest "a pooling multinomial classifier, which provides a platform where composite NB classifier can be launched which, took background knowledge and training examples under consideration". This approach gets the ability to use knowledge from different sources, for text categorization they use two different resources, first the lexical resource and second is the labelled training examples.

This technique presents considerably improved results for different domains as compared to other approaches and hence domain independent. However in between the three blogs used, the political blogs faced the worst accuracy for the following noticed reasons, political posts show more variety, the bloggers compared the candidates and parties, jokes inclusion, anecdotes presence, implicit discussion appropriate with specific candidate, objective and quoted statements from newspapers and more ever the political statements are difficult more to label even for human labelers for the reasons of cultural references for passing judgments , affecting the test set labels.

A lot of more work may be present but we will stress to highlight the approaches more related to the tasks involved in extracting opinion and predict the outcome of some key political event from political tweets.

III. DATA SET

Twitter allowed downloading the real time publically available data with the help of streaming API. Only 1% of the publically available data is allowed to be sampled for free. In Pakistan elections are conducted for local government in the federal capital on the November 30, 2015. In Pakistan elections are party based. The two major parties have been chosen i.e., Pakistan Muslin League Nawaz (PML-N) and Pakistan Tehreek-e-Insaf (PTI). People share their views on the public message board of twitter about the environment in accordance to their observations. We have analyzed 2588 tweets in which there appears any political party name (Pakistan Muslim league Nawaz and Pakistan Tehreek-e-Insaf). These tweets are collected on November 28, on November 29, on the Election Day November 30, 2015. Tweets are searched for specific key words like political party names, their leader names and any other link that clearly pointed out any one of the political party for example BANI GALA is a clear cut nod towards IMRAN KAHN and Pakistan Tehreek-e-Insaf. In the same way captain is another alias for IMRAN KAHN. Tweets containing abbreviations of the party names and other references are also captured. Tweets without these clear cut symptoms are excluded from the tweet volume.

People are actively engaged on twitter to share their ideas about the political parties, their leaders and policies they present and also the future plans of the parties. All those tweets, in which more than two political party names appear, are not included in the analysis. A single user can share only single tweet so the tweets will be normalized and the redundant tweets from a single user are reduced to a single tweet in the data set.

IV. FINDINGS

After the data collection we observed that how many tweets and the number of votes a party got in the polls. We checked the predictive power of twitter that weather twitter has the power of polls prediction.

All the tweets collected are assigned to the sentiment scoring module as mentioned by [17]. The scoring module calculates sentiment score for each of the tweet at word and sentence level. The tweets are classified into subjective tweets and objective tweets. Objective tweets are again pulled out of the race as they are informative and do not convey any sentiment about any political party or their leader. For e.g. objective tweet is" PTI is established in 1996", this tweet only convey an information about the foundation stone of the political party and cannot be used as scale to estimate the user sentiment and affinity towards any political party.

On the other hand subjective tweets fold the user views and interest in the discussion going on so they must be placed in the showcase as they are the building blocks for predicting the outcome of future oriented events. Subjective tweets are further classified into positive, negative and neutral tweets.

The polarized tweets volume is then assigned to filtering module that filters the tweets and categorizes in their respective groups and ties it with a specific political party. Tweets are categorized in their respective groups on the presence of the party names, their leader names or any other link in the tweets. So in this case two groups i.e. (PTI, PMLN) are created and tweets relevant to any one of these are placed under their respective shadow.

The filtering module is characterized by a Named Entity Recognition module which categorizes the polarized tweets. NER is provided with a set of political party names, their leader names, abbreviations and all other prominent references that show some logical link with the political party or party leader or something of the sort. So on the basis of this grouping further calculation will be made. Each party will get positive and negative counts in the tweets. The vote count of the political party will be calculated through the following equations:

$$\text{Vote\_Count (PTI)} = \frac{\mathbf{pos(PTI)+neg(PML-N)}}{\mathbf{pos(PTI)+neg(PTI)+pos(PML-N)+neg(PML-N)}} \quad (1)$$

$$\text{Vote\_Count (PML-N)} = \frac{\mathbf{pos(PTI)+neg(PML-N)}}{\mathbf{pos(PTI)+neg(PTI)+pos(PML-N)+neg(PML-N)}} \quad (2)$$

Twitter have total of 1.9 million users in Pakistan and 10% out of these are from Islamabad [18]. Total of 2588 tweets are collected. The details are given in table 1.

Table: 1. Tweets Details

Tweets	Pakistan Tehreeke-e-Insaf		Pakistan Muslim League Nawaz	
	No of tweets	% share	No of tweets	% share
2 day before	561	55%	459	45%
1 day before	403	50.95%	388	49.05%
Election day	376	48.39%	401	51.61%
Total	1340	51.78%	1248	48.22%

Table 2 shows the polarity classification of the tweets.

Table 2: Tweets Classification

Pakistan Tehreeke-e-Insaf			Pakistan Muslim League Nawaz		
Positive	Negative	Neutral	Positive	Negative	Neutral
209	166	186	187	175	117
164	119	130	145	108	135
137	103	136	168	95	138

Table 3 show confusion matrix which display the results of classified tweets compared with manually annotated set and accuracy of the proposed model.

Table 3: Accuracy of Proposed Model

Tweets	Total tweets	Positive	Negative	Neutral	Accuracy
Positive PTI	146	38	21	87	0.260
Positive PMLN	169	39	38	92	0.230
Negative PTI	96	27	23	46	0.333
Negative PMLN	102	19	35	48	0.343
Neutral	362	99	87	176	0.486
Over all Accuracy					**0.330**

V. CONCLUSION AND FUTURE RESEARCH

The proposed methodology presents an approach for the user contents to uncover the user ideas that is shared on the twitter and used for predicting the outcomes of some key political events like election. Statistics shows that only 10% of Islamabad residents are registered on twitter. So it's not feasible to predict the whole population by this small figure. The gap between predicted and actual results can be minimized for large and more refined samples.

Table: 4 Actual and Predicted Results

Statistics	Pakistan Tehreeke-e-Insaf	Pakistan Muslim League Nawaz
Predicted Seats	25.89	24.11
Actual Seats	17	21

This work can be further elaborated by finding rules for Natural languages which is trivial. Finding patterns for a human language that can be reused and executed by computers is appropriate. Opinion mining can also be enhanced to understand the semantics of texts in more intelligent way. Enhancing the accuracy of the predictive model and reduction of the human efforts in the analysis leads to an interesting direction for future research.

REFERENCES

[1] Woodly, D. 2007. New competencies in democratic communication? Blogs, agenda setting and political participation. *Public* Choice, 134(1-2): 109-123.

[2] Sunstein, Cass. 2007. Neither Hayek nor Habermas. *Public Choice*, 134(1-2): 87-95.

[3] Jansen, B. J.; Zhang, M.; Sobel, K.; and Chowdury, A. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60: 1 20.

[4] Jansen, HJ, and Koop. R. 2005. Pundits, Ideologues, and Ranters: The British Columbia Election Online. *Canadian Journal of Communication*, 30(4): 613-632.

[5] Schneider, S. M. 1996. Creating a Democratic Public Sphere Through Political Discussion. *Social Science, Computer Review*, 14(3): 373-392.

[6] Farrell, H., and Drezner D.W. 2008. The power and politics of blogs. *Public Choice*. 134(1-2): 15-30.

[7] Williams, C., and Gulati, G. 2008. What is a Social Network Worth? Facebook and Vote Share in the 2008 Presidential Primaries. In *Annual Meeting of the American Political Science Association*, 1-17. Boston,MA.

[8] Skemp, K. 2009. All A-Twitter about the Massachusetts Senate Primary. Retrieved December 15, 2009 from http://bostonist.com/2009/12/01/ massachusetts senate primary debate twitter.php

[9] Véronis, J. 2007. Citations dans la presse et résultats du premier tour de la présidentielle 2007. Retrieved December 15, 2009 from http://aixtal.blogspot.com/ 2007/04/2007-la-presse-fait-mieux-que-les.html

[10] Adamic, L. A., and Glance, N. 2005. The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, 36-43. Chicago, IL.

[11] Cozma, Raluca, and K. Chen. "Congressional Candidates' Use of Twitter During the 2010 Midterm Elections: A Wasted Opportunity?." *61st Annual Conference of the International communication association*. 2011.

[12] Pew Research Center, "Parsing Election Day Media: How the Midterms Message Varied by Platform", Pew, 2010, retrieved June 14, 2011, from http://pewresearch.org/pubs/1794/parsing-election-daymedia- messages-varied-by-platform

[13] Romero, D. M., Meeder, B., & Kleinberg, J. (2011, March). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World Wide Web* (pp. 695-704). ACM.

[14] Bollen, J., Mao, H., and Pepe, A., "Modeling Public Mood and Emotion: Twitter Sentiment and Socioeconomic Phenomena", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011), 2011, retrieved August 30, 2011, from http://arxiv.org/abs/0911.1583v1

[15] Durant, Kathleen T., and Michael D. Smith. "Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection." *Advances in Web Mining and Web Usage Analysis*. Springer Berlin Heidelberg, 2007. 187-206.

[16] Melville, Prem, Wojciech Gryc, and Richard D. Lawrence. "Sentiment analysis of blogs by combining lexical knowledge with text classification." *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009.

[17] Asghar, MZ. "Lexicon based Approach for Sentiment Classification of User Reviews." *Life Science Journal* 11.10 (2014): 468-473

[18] Twitter users in%C2%A0Pakistan.html