International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 5, Issue 2, March - April 2016 **ISSN 2278-6856**

Predicting Bihar Vidhan Sabha Elections Results with Twitter Using Bayesian Classifier

Varsha D. Jadhav[1], Sachin N. Deshmukh[2]

[1]P.E.S. College of Engineering, Aurangabad, Maharashtra, India

[2]Department of Computer Science and Information Technology, Dr, Babasaheb Ambedkar Marathwada University Aurangabad,Maharashtra India

Abstract

The data on social networking sites expresses opinions of people about business organizations, customer reviews and many real world applications. It is now a common practice to use social media data for political analysis, especially during election time. We focus our work on using election tweets from Twitter from the districts of Bihar. The objective of this paper is to predict public intention for the results of Bihar Vidhan Sabha election held in October-November 2015. This work uses the Bayesian method in R as the technology. This is helpful for predicting the results on the basis of public opinion given on the Twitter.

Keywords: Twitter, Sentiment, Polarity, Emotion, Intention, Polarity, Bayes.

1. INTRODUCTION

Today the people are connected using Internet. Social sites are specially used by young youths for connecting the world because they spend lot of the time on sites like Facebook, Twitter etc. Every current topic on social media is supported by public opinion which makes clear the intention of that particular topic to the world. In this paper we center our work on Twitter which is one of the popular micro blogging sites which connect billion of people on a single platform. We are doing analysis of data on Twitter related to Bihar elections held in October-November 2015. The election was between Grand Alliance and National Democratic Alliance (NDA).Grand Alliance was formed by JDU, RJD and Congress. Whereas BJP, LJP, RLSP and HAM formed the National Democratic Alliance. There was strong competition between Grand Alliance and National Democratic Alliance. Here we were interested in extracting the information related to Grand Alliance and NDA which were the main alliances formed for these elections. To extract this information R technology is used. The election related tweets were downloaded using the keywords as Grand Alliance Bihar and NDA Bihar. In

India majority of the people are from urban region, so the tweets from districts are considered. Bayesian classifier was used to predict the emotions and sentiments related to the extracted Bihar election tweets.

2. LITERATURE REVIEW

An emerging area of research is result forecasting, such as predicting the election results, which is the intention of this paper. Andranik Tumasjan et al. [1] showed that Twitter is indeed used extensively for political deliberation. Adam Bermingham et al. [2] concluded that Twitter does appear to display a predictive quality which is marginally augmented by the inclusion of sentiment analysis. Felipe Bravo-Marquez et al. [3] analyzed that Twitter extracted opinion time series related to a specific event are or not suited for generating predictive models. Erik Tjong Kim Sang et al. [4] performed prediction of 2011 Dutch Senate Election results with Twitter by counting the tweets that mention political parties and showed that it is not sufficient to obtain good predictions. Zhongyu Wei et al. [5] have studied media behavior on Twitter during the UK's General Election in 2010 and showed that while most information flows are originated from media, they seem to lose their dominant position in shaping public opinion during the UK's general election. Mark Huberty [6] showed that simplistic methods for forecasting elections from Twitter, even when their results are correlated with election outcomes, provide relatively little added benefit. Priya Sharma et al. [7] predicted the result of Delhi election held in February 2015 by analyzing the tweets using Hadoop as technology. Andreas Jungherr [8] offers a detailed analysis of Twitter messages posted during the run-up to the 2009 federal election in Germany and their relationship to the electoral fortunes of Germany's parties and candidates, which focus for measuring the attention on parties and candidates on Twitter and the relationship to their respective vote share. Jasmina Smailovi [9] monitored the Twitter sentiment during the Bulgarian elections.

3 DATA COLLECTION AND TRAINING THE CLASSIFIER

3.1 Data Collection

Data was collected using Twitter API from 6th November to 8th November 2015 for the keywords Grand Alliance Bihar, NDA Bihar. The collected dataset is used to extract features that will be used to train the sentiment classifier. Experimentation is carried out using n-gram binary features.

i) Filtering: Remove URL links, Twitter user names, Twitter special words such as RT.
ii) Tokenization: segment text by splitting it by spaces and punctuation marks, and form a bag of words.
iii) Removing stopwords: we remove articles ("a", "an", "the") from the bag of words.
iv) Removing punctuation, numbers and unnecessary spaces.
v) Converting to lower case: All the letters in the sentences are converted into lower case
vi) Constructing n-grams: we make a set of n-grams out of consecutive words.

3.1 Bayes Classifier

Sentiment classifier is build using the Naive Bayes classifier. Naive Bayes classifier is based on Bayes theorem.

$$P(s \mid M) = \frac{P(s) \bullet P(M \mid s)}{P(M)} \qquad (1)$$

Where:
s is a sentiment, M is a Twitter message.

Two Bayes classifiers are trained, which use different features: presence of n-grams and part-of-speech distribution information. N-gram based classifier uses the presence of an n-gram in the post as a binary feature. The classifier based on POS distribution estimates probability of POS-tags presence within different sets of texts and uses it to calculate posterior probability. Although, POS is dependent on the n-grams, an assumption of conditional independence of n-gram features and POS information for the calculation simplicity is made

$$P(s \mid M) \approx P(G \mid s).P(T \mid s) \qquad (2)$$

Where:
G is a set of n-grams representing the message;
T is a set of POS-tags of the message.
We assume that n-grams are conditionally independent:

$$P(G \mid s) = \prod_{g \in G} P(g \mid s) \qquad (3)$$

Similarly, we assume that POS-tags are conditionally independent: (3)

$$P(T \mid s) = \prod_{t \in G} P(t \mid s) \qquad (4)$$

$$P(s \mid M) \approx \prod_{g \in G} P(g \mid s) \bullet \prod_{t \in G} P(t \mid s) \qquad (5)$$

Finally, we calculate log-likelihood of each sentiment:

$$L(s \mid M) = \sum_{g \in G} \log(P(g \mid s)) + \sum_{t \in G} \log(P(t \mid s))$$

4 SENTIMENT ANALYSIS IN R

Two packages in R that can be used for sentiment analysis are *sentiment* and *qdap*. In the experimentation *sentiment* package is used. This package requires *tm* and *Rstem* packages, so they should be installed first. The function classify_polarity is called which retrieve the column corresponding to best-fit.
Sentiment analysis techniques can be classified into two high level categories:

1. **Lexicon based:** This technique work on an assumption that the collective polarity of a sentence is the sum of polarities of the individual words or phrases.
2. **Learning based:** These techniques require training a classifier with examples of known polarity presented as text classified into positive, negative and neutral classes.

R's sentiment package follows a lexicon based approach. In the R package library under \sentiment\data\data folder the lexicon is found as a file named subjectivity.csv.g

Results and discussion

For the collected data for the Bihar Vidhan Sabha elections 2015 polarity was classified. Assigning polarity is classifying the tweets as positive, negative and neutral. The accuracy measures such as Mean error (ME), Root mean square error (RMSE), Mean absolute error (MAE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE). MAE is simply the mean of the absolute errors. The absolute error is the absolute value of the difference between the forecasted value and the actual value. MAE tells us how big of an error we can expect from the forecast on average. Cort J. Willmott et.al [10] indicates that MAE is the most natural measure of average error magnitude than RMSE.There was a strong fight between National Democratic Alliance and Grand Alliance. We analyze the accuracy using the mean absolute error (MAE) on each day which is shown in Table 1.

Table 1: Analysis of NDA and Grand Alliance (GA) based on MAE.

	6th Nov 2015	7th Nov 2015	7th Nov 2015
NDA	3.617458	2.985608	2.985608
Grand Alliance	3.544254	3.897907	3.901159

On 6th November 2015 Grand Alliance have MAE less than NDA with an accuracy of 96.455746%.On 7th and 8th November 2015 MAE for NDA is less than Grand Alliance with an accuracy of 97.014392%. On 6th November 2015 Grand Alliance is leading NDA, whereas on 7th and 8th November NDA is leading Grand Alliance.
There are 38 districts in Bihar. People mostly from urban region tweet, so results of urban region are considered. From the 38 districts, data of Rohtas and Saran districts was not available, so data of 36 districts was considered.JDU, RJD, and Congress together form Grand Alliance. JDU won 6 seats, RDJ won 9 seats and Congress

won 7 seats. NDA is formed by BJP, LJP, RLSP and HAM. In the districts region LJP, RLSP and HAM did not win any seat, whereas BJP alone won 14 seats. When considered single party, BJP alone had won 14 seats which are more than the seats won by JDU, RJD and Congress. BJP strongly supports NDA and so more people tweeted positively for NDA to be in power which nearly satisfies our results.

In figure 1 tweet between NDA and Grand Alliance are shown. The tweets are plotted against dates. It shows the comparative analysis of two party's tweets. This type of analytics predicts the negative, neutral and positive tweets.

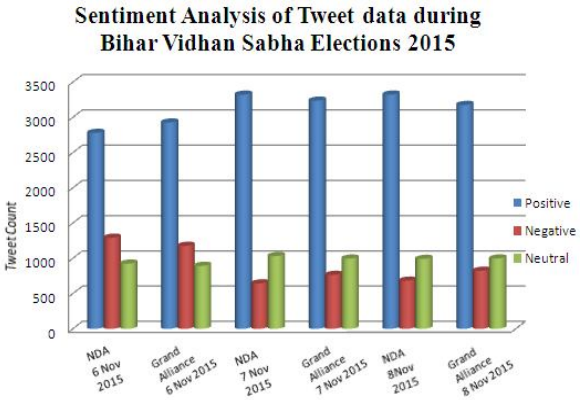


Figure 1 Analysis of tweets between NDA and Grand Alliance

References

[1] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

[2] Adam Bermingham and Alan F. Smeaton, 2011, "On Using twitter to Monitor Political Sentiment and Predict Election Results," In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011, pages 2-10, Chiang Mai, Thailand.

[3] Felipe Bravo-Marquez, Daniel Gayo-Avello, Marcelo Mendoza and Barbara Poblete, 2012, "Opinion Dynamics of Elections in Twitter," In IEEE eighth Latin American Web Congress 2012.

[4] Erik Tjong Kim Sang, Johan Bos, 2012, Predicting the 2011 Dutch Senate Election Results with Twitter, In: Proceedings of SASN 2012, the EACL 2012Workshop on Semantic Analysis in Social Networks, Avignon, France.

[5] Zhongyu Wei, Yulan He, Wei Gao, Binyang Li, Lanjun Zhou, Kam-fai Wong, 2013, Mainstream Media Behavior Analysis on Twitter: A Case Study on UK General Election, In 24th ACM Conference on Hypertext and Social Media 1–3 May 2013, Paris, France.

[6] Mark Huberty, 2013, Multi-cycle forecasting of Congressional elections with social media., In CIKM'13, , San Francisco, CA, USA.

[7] Priya Sharma,Pankaj Kumar Dwivedi, 2015, Big-Data Analytics To Predict Election Results, International Journal of Advanced Research in Computer Science & Technology (IJARCST 2015), Vol. 3, Issue 2.

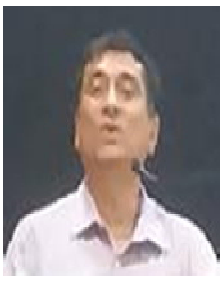
[8] Andreas Jungherr, 2013, Tweets and Votes, a Special Relationship. *PLEAD'13*, San Francisco, CA, USA.

[9] Jasmina Smailovic, Janez Kranjc, Miha Grcar, Martin, Znidarsic, Igor Mozetic, 2015, Monitoring the Twitter sentiment during the Bulgarian elections. In IEEE.

[10] Cort J. Willmott, Kenji Matsuura, 2005, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model Performance, Climate Research, Vol. 30: 79–82.

AUTHOR(4)

Varsha D.Jadhav is currently working as an Assistant Professor in Department of Computer Science and Engineering of P.E.S. College of Engineering Aurangabad, and pursuing PhD in Computer Engineering from Dr Babasaheb Ambedkar Marathwada University, Aurangabad. Research area is Data mining in Social Networking. She published research papers in various international conferences and journals



Dr. Sachin N. Deshmukh is Currently working as Professor in Department of Computer Science and IT, Dr Babasaheb Ambedkar Marathwada University, Aurangabad and having experience of around twenty years in teaching for post graduate (M. Tech, M.Sc. and MCA) and graduate courses (B.E., B. Tech) University Authorities also have given the responsibility as Director (University Network Information Center), Director (Center for Vocational Education and Training). Area of research is Text mining, Social Web mining and Intension Mining and recently has completed Research Project on this funded by AICTE. He also worked on research projects of UGC and AICTE. He has published 40 research papers in various international journals and conferences.