Driving transportation policy using social media data: Case study on the Delhi Odd-Even policy

Pranamesh Chakraborty[1] , and Anuj Sharma[2]

[1]Ph.D. Student, Department of Civil Construction and Environmental Engineering, Iowa State University, Ames, Iowa-50010, USA; E-mail: pranames@iastate.edu

[2]Associate Professor, Department of Civil Construction and Environmental Engineering, Iowa State University, Ames, Iowa-50010, USA; E-mail: anujs@iastate.edu

Abstract. Twitter, a microblogging service, has become a popular platform for people to express their views and opinions on different issues. Sentiment analysis of tweets can help in understanding public opinion on different government decisions. This paper uses Twitter data to extract sentiments of people during the Phase 1 and Phase 2 of the odd-even policy implemented by the Delhi government to curb the air pollution and improve traffic flow. Sentiment analysis can be done using lexicon-based approach or text classification based approach. In this study, we used four different lexicon based approaches: Bing, Afinn, NRC and CoreNLP to extract sentiments from tweets and thereby assess overall public opinions. The daily trend obtained for each phase are normalized with the number of tweets and then compared using Granger causality test. The causality test results show that the trends obtained during the two phases were different from each other and no causal relationship were found between the trends of the two phases. In particular, public sentiments mostly turned negative during the later stage of the Phase 2. Such analyses are expected to help the policy makers provide a perspective of public opinion towards similar transportation policies and thereby guide their future movements.

Introduction

Delhi and its National Capital Region (NCR), with a population of 25.8 million constitutes 7.8% of India's urban population [1]. A combination of 200 km of metro rail, buses and different para transit options (auto rickshaws) serve the public transportation of Delhi. Although the share of cars in community trips in Delhi is relatively low compared to other major cities of the world, recent studies have ranked Delhi as the "worst" polluted city in terms of environment performance index [2]. To tackle the high pollution level, the Government of the National Capital Territory of Delhi (GNCTD), India, implemented the odd-even policy, where only odd numbered cars were allowed to operate on odd numbered dates and cars with even number plates on even days between 08:00-20:00 hours. This policy was implemented in two phases. The first phase was implemented in winter season, between January 1 and 15, 2016 and the second phase during summer from April 15 to April 30, 2016. Twenty different categories were exempted from this rule which included motorized two-wheelers, electric and hybrid cars, cars driven by women, cars of very/very important persons such as parliamentarians, emergency vehicles such as ambulance, fire brigade, etc.

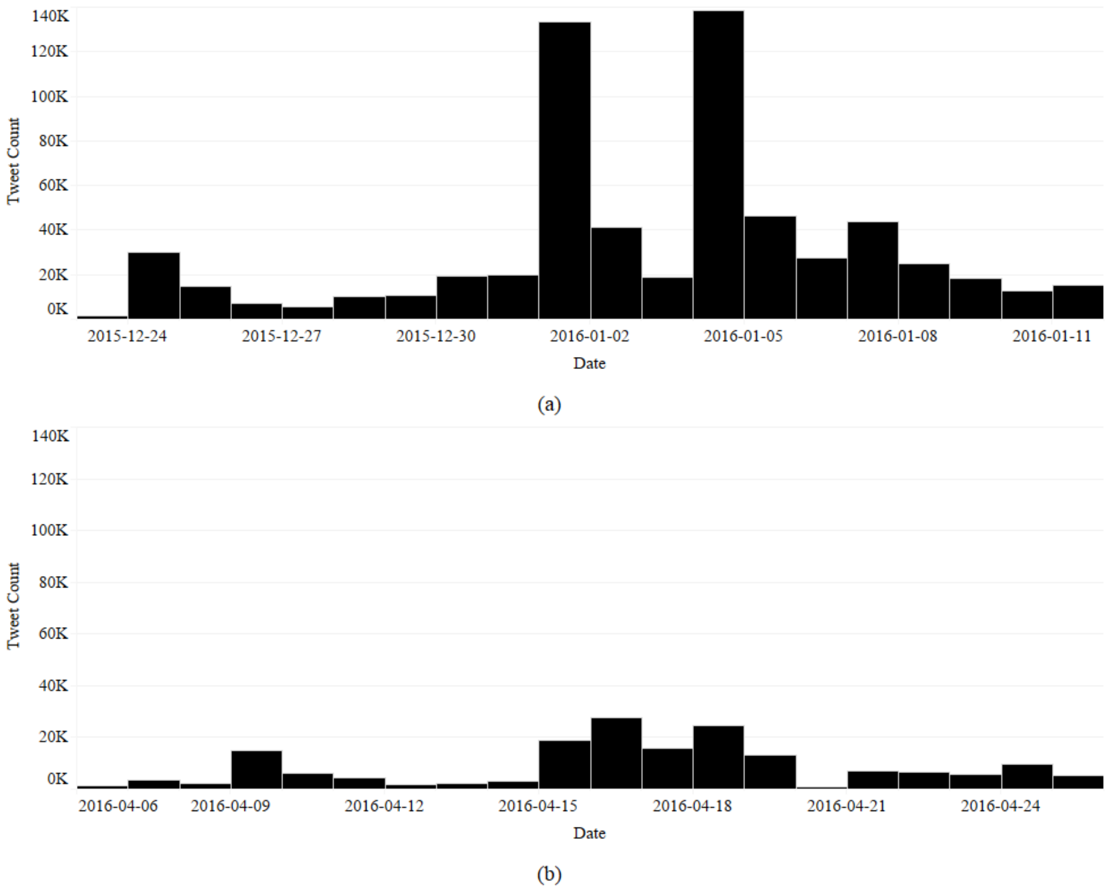
Similar such odd-even policy have been implemented in the past in other parts of the world too. This includes Buenos Aires (Argentina), Bogota (Columbia), Mexico City (Mexico), Manila (Philippines), Lagos (Nigeria), and Beijing (China) during the Olympic Games [3]. However, past studies suggest that although such driving restriction policies may reduce pollution and congestion in short run, but in long run people learn to cope with the restrictions by shifting to two-wheelers, buying second older cars, etc. [4]. Hence, studies have been done to find out the efficacy of the given implementation.

Chelani (2017) [5] analyzed the concentration of PM2.5 during the odd-even policy dates. Similarity and causality analysis were performed to find out the local and regional influence. Kumar et al. (2017) [6] also conducted similar study to find out the influence of the policy on fine and coarse particles. The study suggested that even though the certain hours of the trial day generated cleaner air, but the overnight emissions from heavy goods vehicles made them overall ineffective. Mohan et al. (2017) [7] evaluated the car, two-wheeler, bus and auto rickshaw flow rates and also car occupancy rates during the traffic restriction experiment period. They found that the car flow rates decreased by 20% while other modes of vehicles increased during the analyses period. However, no significant increase in car occupancy rates were observed during the period which suggests that the car owners didn't opted for car sharing. Although significant amount of research have been performed to evaluate the effects of the odd-even policy on the environmental and traffic aspect, research can also be performed to find out the public opinion towards such a policy implementation. Social media platforms (e.g. Twitter and Facebook) can be used to extract such information and mine them to get useful information.

Social media has gradually evolved to be a popular platform for people to express their opinions on current trending topics. Twitter is such a platform where users can post brief text updates (maximum 140 characters) or multimedia such as images or audio clips. Researchers are using the "tweets" to find out general public perceptions on a variety of topics [8]. These sources have been used for monitoring political sentiments, predicting election results [9], detecting tension in online communities [10], understanding sentiments on new product launch [11], etc. Besides these, Twitter has also been used for tracking complex real-time events like natural disasters [12], [13], road hazards detection [14], and disease propagation [15]. Collins et al. (2012) [16] used Twitter data for monitoring sentiments of the riders of the Chicago Transit Authority public transit system. Luong et al. (2015) [17] also conducted a similar study to find out public opinion of the light rail service in Los Angeles. Sasaki et al. (2012) [18] performed a feasibility study using Twitter as a sensor for detecting transportation information. Recently, Sharma et al. (2017) [19] used Deep Belief Network (DBN) to classify the sentiments of tweets posted by users during the odd-even policy in Delhi. They used six different models based on the DBN classifier to find out the performance of the proposed methods. This paper on the other hand uses four different lexicon-based approaches to perform the sentiment analysis of the tweets collected during the 1st and 2nd phase of the odd-even policy implementation. The accuracy of the methods are evaluated based on 500 randomly sampled tweets from the given dataset. Then, the overall daily trend of the sentiments during the two phases obtained from the different methods are compared based on Granger causality test to check their similarity/dissimilarity during the two phases. Such a trend comparison can help policy makers to understand the perspective of public opinion towards such transportation policies and thereby guide their future movements.

Data Description

The twitter data for the 1st phase of the policy implementation were bought using the Full Archive Search API provided by GNIP. For the 2nd phase of the policy implementation, Twitter Streaming API was used for collecting tweets in real-time related to the Odd-Even policy. The popular hashtags (e.g., #OddEven, #oddevenformula, #oddevenplan, #OddEvenRule, #OddEvenPolicy, #oddevendobara) were used for extracting the relevant tweets for the 1st and 2nd phase of the policy implementation. Relevant tweets were downloaded from 9 days before the policy implementation to the first 11 days of the policy implementation resulting in a total of 20 days for each phase of the policy implementation. A total of 650,000 tweets were obtained during the 1st phase from 23rd December, 2015 to 11th January, 2016 while 180,000 tweets were obtained during the 2nd phase from 6th April, 2016 to 25th April, 2016. Using the Full Archive Search API (Enterprise version) for downloading the tweets during the 1st phase of the policy resulted in larger number of tweets compared to the tweets of the 2nd phase obtained using the free Twitter Streaming API. Figure 1 shows the number of tweets collected during each day of the analysis period. It can be seen that the volume of tweets increased during the initial period of the policy implementation (around 1st January and 15th April) for the both the phases. Retweets were also included in the data source with the assumption that if a user retweets, it indicates that the user is also having a similar opinion or sentiment.



(a)

(b)

Figure 1. Daily number of tweets downloaded during the (a) Phase 1 and (b) Phase 2

Pre-processing

The twitter data often contains noise such as RT for retweets, URLs, @usernames, etc. which needs to be removed before the sentiment analysis task. Our preprocessing step involved the standard preprocessing steps used in previous literature [20], [21] which includes removal of (a) URLs or external website links, (b) hashtags, (c) stop words such as ‘a’, ‘an’, ‘the’, etc., (d) usernames, (e) unnecessary spaces, (f) punctuation marks, (g) numbers, and (h) special characters such as emotions or non-English alphabets such as Hindi, etc. Finally, the stemming process is applied to convert all inflected words to its root form called ‘stem’. For example, ‘automatic’, ‘automation’, and ‘automate’ are converted to its stem form ‘automate’. Snowball stemmer, the popular stemming package is used for this purpose.

Methodology

Sentiment analysis of tweets involves determination of the polarity of the tweets, whether it is expressing positive, negative or neutral sentiment towards the topic/subject. Hence sentiment classification can be also termed as polarity determination. Four different classes of twitter sentiment analysis approaches have been identified in the literature [21]

- Machine learning
- Lexicon based
- Hybrid (Machine learning & Lexicon based)
- Graph based

A majority of machine-learning methods involves building a classifier from machine-learning domain trained on different features to detect sentiment of tweets. The common classifiers used are Support Vector Machines (SVM), Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Conditional Random Field (CRF). Details of such application can be found in [22], [23]. Lexicon based methods, on other hand, use a pre-determined list of positive and negative terms to determine the polarity of the tweets. Details of these methods will be discussed later in this section. Hybrid approach methods combine lexicon-based and machine-learning methods while graph-based methods use social network properties to achieve better performance [24], [25]. This study uses four different lexicon-based methods to determine overall public opinion on the odd-even policy implemented by the Delhi government. Details of the methods used in this study are discussed next.

Different lexicon-based methods exist which utilize different sets of pre-determined list of opinion words to determine the overall polarity of the tweet or text. This paper uses four such lexicon-based methods namely (a) Afinn (b) NRC (c) Bing and (d) Stanford CoreNLP to determine the sentiments of the extracted tweets. The Afinn lexicon [26] includes 2,477 English words with positive words scored from 1 to 5 and negative words from -1 to -5. The word list is focused on language commonly used in microblogging platforms like Twitter and hence contains acronyms, web jargons and slang words too. The NRC lexicon [27] is also a similar word-emotion association lexicon containing more than 14,000 distinct words created by using the crowdsourcing Amazon Mechanical Turk. The Bing [28] lexicon is also a similar lexicon dictionary containing around 6800 positive and negative English opinion words. All these words assigns points for each positive and negative words in a tweet and then sum up these points to find out the overall sentiment of the tweet. The Stanford CoreNLP method [29], on other hand, not only uses the positive and negative words, but also utilizes the order of the words to build the overall polarity of the sentiment. The model is based on Recursive Deep Neural Network that builds on top of grammatical structures.

Evaluation of sentiment classification

Each of these four sentiment classification methods are applied on the extracted tweets to find out the sentiments. Five hundred randomly sampled tweets are selected and hand-annotated into three classes: positive, negative and neutral. Precision and recall (Equations 1 and 2 respectively) are computed for each class and then the average precision and recall are determined. Finally, the F-measure (Equation 3) is also computed.

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{1}$$

$$decision = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2}$$

$$F = 2.\frac{precision.recall}{precision + recall} \tag{3}$$

Daily trend comparison

The sentiment value assigned to each tweet by the given lexicon-based methods is determined and aggregated to get the overall sentiment of each day, denoted as SS_d^m, where *d* denoted the date and m denotes the method. Since the scale of sentiment scores given by each method is different, SS_d^m is normalized based on the range of sentiment score (RS^m) given by method m. RS^m is obtained by finding out the absolute difference of maximum and minimum sentiment score to the analyzed tweets of the dataset by the method *m*. Also, to account for the varying number of tweets obtained for each day, SS_d^m is also normalized by the daily number of tweets (n_d). The daily normalized sentiment strength, denoted by NSS_d^m, is given by Equation 4. The variation of NSS_d^mdepicts the daily trend of public opinion on the Odd-Even policy.

$$Normalized\ Sentiment\ Strength\ (NSS_d^m) = \frac{SS_d^m}{n_d \times RS^m} \tag{4}$$

The daily trend of the sentiment score obtained by each method are then compared to find out the overall agreement of the trend given by each method. To compare the trend, Granger causality test is performed.

Granger causality test checks that if predictions of a variable *Y* can be improved by using its own past values and also past values of another variable *X* rather than using its own past values only. Let *Y* follows a univariate linear autoregressive models given by Equation 5. Then, the autoregression is augmented including the lagged values of the variable X, given by Equation 6. Granger test can be performed by an F-test which reports the Walds statistics for the joint hypothesis given in Equation 7. Rejection of the null hypothesis can be inferred as X Granger-causes Y.

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + +\alpha_m y_{t-m} + \varepsilon_t \tag{5}$$

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + +\alpha_m y_{t-m} + \beta_1 x_{t-1} + \cdots + \beta_l x_{t-l} + \varepsilon_t \tag{6}$$

$$\beta_1 = \beta_2 = \cdots = \beta_l \tag{7}$$

Results

Table 1 shows five sample tweets and the sentiment scores assigned by each method. Since most of these methods (except CoreNLP) do not take into account the sentence structure and rather look at words in isolation, giving positive points for positive words and negative for negative words, hence they fail to classify correctly complex structured tweets (e.g., Tweet # 5 in Table 1).

Table 1. Sentiment score assigned by the different methods

#	Tweet	Sentiment Score assigned by method			
		Bing	Afinn	NRC	CoreNLP
1	Another vile attempt to make #OddEven a failure, this time with a dangerous excuse. #BJP proxy #VHP in action	-4	-6	-2	-5
2	#OddEven is a gimmick apart from reducing vehicles on roads. Kejri is burning tax money on publicity and people keep choking on diesel fumes	-3	-2	-0.9	-1
3	Delhi #oddeven: @Olacabs, @Uber surge pricing puts commuters in a jam, reports @mallicajoshi	-1	0	1	0
4	Smooth ride today #OddEven #OddEvenDobara less of cars only odd ones showing up... Good show #delhi	1	1	1	-1
5	Even in terms of traffic, #OddEven doesn't seem to be as successful as last time.	1	3	1	-3
6	"RT @IndiaToday: #OddEvenPlan was supported not just by Delhi people but even, judges car-pooled and walked to work: #Kejriwal	2	2	1	-1

In order to find out overall accuracy of these methods, 500 randomly sampled tweets were hand-annotated and compared with the output obtained from the algorithms. Table 2 shows the precision, recall and F-score (Equations 1, 2 and 3) of each method. As can be seen from Table 2, the accuracy and other metrics for CoreNLP are substantially lower compared to the other methods. One reason can be the CoreNLP algorithm presently is designed for analysing sentiments of proper English sentences rather than tweets which generally consist of a significant amount of web jargon and hence do not perform well in this case. Due to poor performance of CoreNLP in the test dataset, it has been excluded from further analysis and the remaining of the analyses have been performed using three other algorithms (Bing, Afinn and NRC).

Table 2. Precision, Recall and Accuracy of each method used

Method	Precision	Recall	F-score
Bing	0.688	0.683	0.685
Afinn	0.703	0.673	0.688
NRC	0.632	0.760	0.690
CoreNLP	0.483	0.388	0.430

Daily trend estimation and comparison

The sentiments obtained from the tweets were aggregated to get the daily trend of the normalized sentiment strength, *NSS* (Equation 4). To recall, normalization is performed to take into account the daily variation of the number of tweets and also scale of sentiment scores assigned by each method. To compare the trends during the two phases of the policy implementation, relative dates are used with Day 1 being the start date of the policy implementation (i.e., 1st January and 15th April respectively). Figure 2 shows the trend of *NSS* obtained from each method during the analysis period. It shows that even though people were enthusiastic during the initial period of the Phase 2 (April 15th), however the sentiment scores kept decreasing during the later phase of the policy implementation. On the other hand, sentiment scores were steady and mostly positive even during the later stage of the Phase 1. To compare the similarity of the trends obtained from each method, Granger causality test is performed. Table 2 gives the *p*-value obtained for each method pair. To recall, the null hypothesis states that there is no Granger causality between the two time-series. Hence, higher p-values (> 0.05) given in Table 2 proves that we fail to reject the null hypothesis that there is no Granger causality between the trend of sentiments obtained during the two phases. In other words, there is significant difference between the trends obtained from the two phases. This is also evident from Figure 2 which shows the sentiments scores dropped during the later stage of the 2nd phase of the policy implementation.

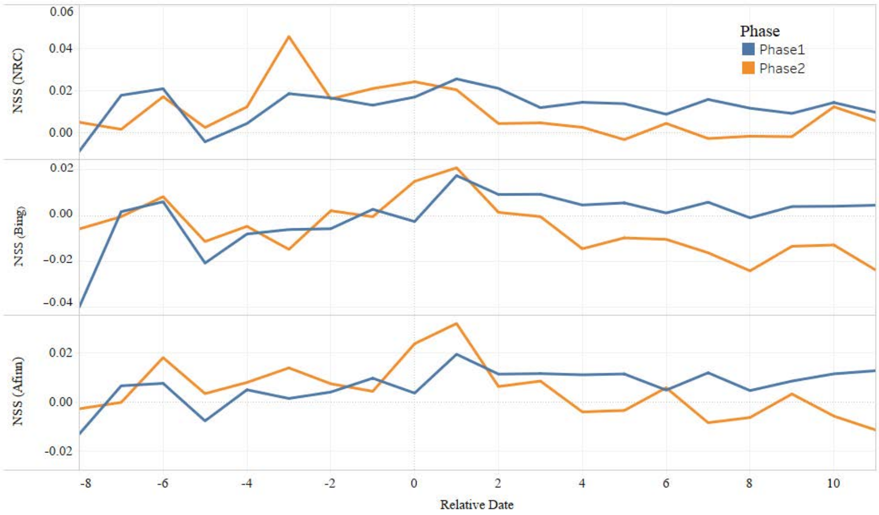


Figure 2. Daily trend of the normalized sentiment strength (NSS) using different methods during the two phases of policy implementation

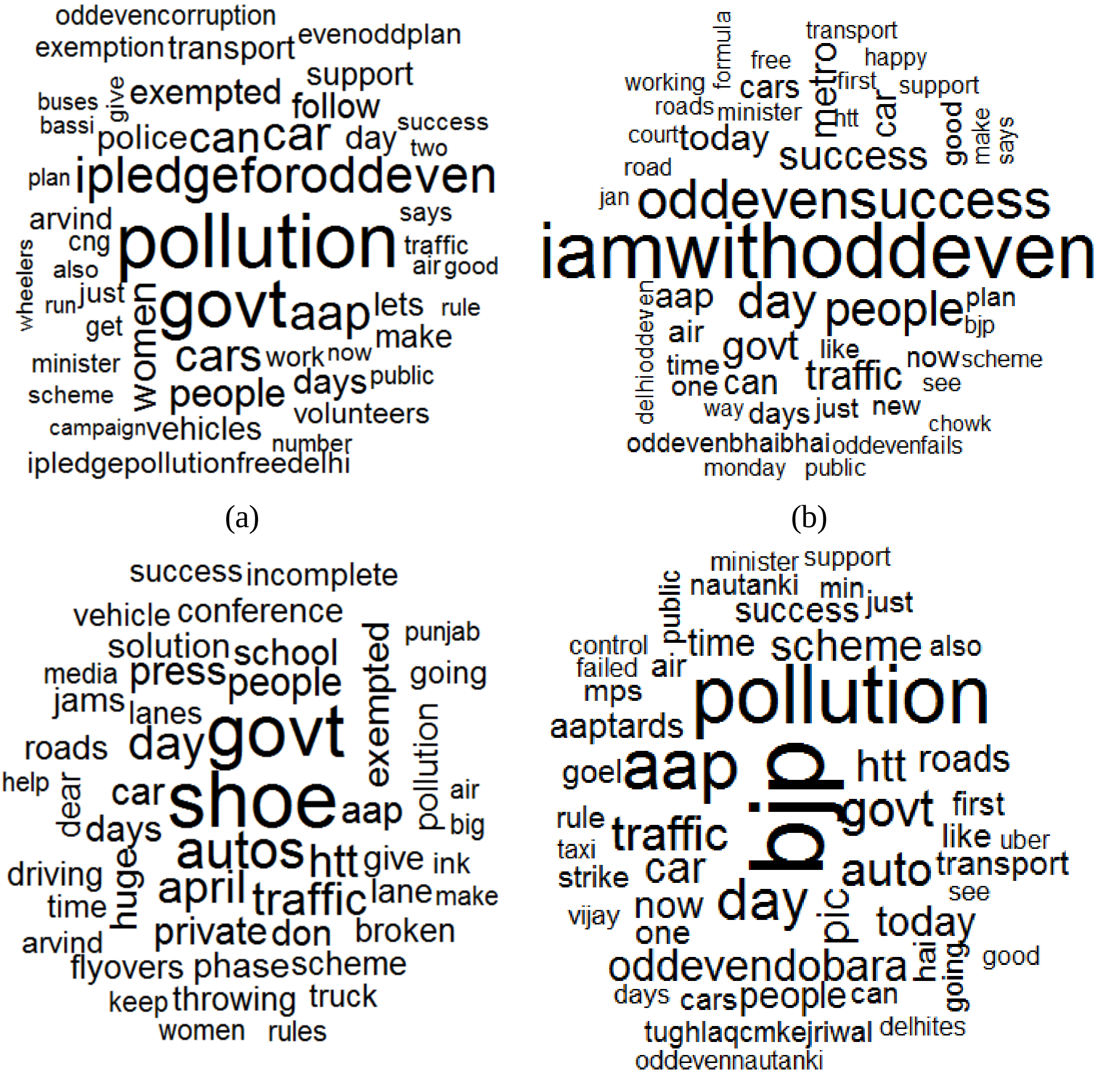
Table 3. Daily trend comparison by Granger test

Method used	***p*-value**
NRC	0.435
Bing	0.178
Afinn	0.446

Wordclouds are also plotted to find out the frequently occurring words during the two phases. In order to find out the sentiments during the pre-phase policy implementation separately, the periods before the policy implementation are now classified as separate groups: Pre-Phase 1 (23rd December, 2015 to 31st December, 2015) and Pre-Phase 2 (6th April, 2016 to 14th April, 2016). Phase 1 and Phase 2 are now defined from 1st January, 2016 to 11th January, 2016 and 15th April, 2016 to 25th April, 2016. Figure 3 shows the wordclouds during the Pre-Phase 1, Phase 1, Pre-Phase 2, and Phase 2 respectively. Similar to Figure 2, Figure 3 also shows that positive comments like “Iamwithoddeven”, “oddevensuccess” which were frequent during Phase 1 were not present during Phase 2 suggesting the negative sentiments of public during the Phase 2 implementation.

Conclusions

Social media has gradually evolved to be a popular platform for people to express their views on different topics. These resources can be used for monitoring public sentiments during different trending issues, new product launch, etc. This study uses Twitter data to



(a) (b)

(c) (d)

Figure 3. Wordcloud of tweets during (a) Pre-Phase 1 (b) Phase 1, (c) Pre-Phase 2, and (d) Phase 2 of Odd-Even policy

find out public sentiments during the Odd-Even policy implemented in Delhi, the national capital of India. Four different lexicon-based approaches have been used for sentiment analysis purpose. Accuracy of these methods have been determined on hand-annotated test set. And finally, the daily trend of the sentiment during the two phases of the policy implementation are determined. Causality tests are performed to check the similarity of the trends obtained during the two phases. The overall trend shows that although people were enthusiastic during the initial period of the 2nd phase of the policy implementation, however, more people started giving negative views on the policy with the progress of the Phase 2. Such trend analysis can help the government to find out the perspective of public opinion towards different policies that can help then in guiding their future movements. In future, detailed study can be performed to find out the grievances in the public and what measures can be adopted to tackle such issues. Also, more advanced machine learning algorithms (e.g., Support vector machines, etc.) can be used in future to obtain better results on the sentiment analysis of the tweets.

Acknowledgement

The authors gratefully acknowledge the help provided by Aditya Raj and Abhinav Prakash in downloading Twitter dataset used in this study.

References

[1] N. D. Office of the Registrar General and Census Commissioner, “Census of India. Provisional population totals 2011,” 2012 .

[2] A. Hsu and A. Zomer, “An Interactive Air-pollution Map,” 2014.

[3] L. Wang, J. Xu, and P. Qin, “Will a driving restriction policy reduce car trips?-The case study of Beijing, China,” *Transp. Res. Part A Policy Pract.*, vol. 67, pp. 279–290, 2014.

[4] F. Gallego, J.-P. Montero, and C. Salas, “The effect of transport policies on car use: Evidence from Latin American cities,” *J. Public Econ.*, vol. 107, June, pp. 47–62, 2013.

[5] A. B. Chelani, “Study of Local and Regional Influence on PM2.5 Concentration during Odd-Even Rule in Delhi Using Causal Analysis,” *Aerosol Air Qual. Res.*, vol. 17, no. 5, pp. 1190–1203, 2017.

[6] P. Kumar, S. Gulia, R. M. Harrison, and M. Khare, “The influence of odd–even car trial on fine and coarse particles in Delhi,” *Environ. Pollut.*, vol. 225, pp. 20–30, 2017.

[7] D. Mohan, G. Tiwari, R. Goel, and P. Lakhar, “Evaluation of Odd-Even Day Traffic Restriction Experiments in Delhi, India,” in *TRB 96th Annual Meeting Compendium of Papers*, 2017, November 2016, pp. 1–14.

[8] M. Zimmer and N. J. Proferes, “A topology of Twitter research: disciplines, methods, and ethics,” *Aslib J. Inf. Manag.*, vol. 66, no. 3, pp. 250–261, 2014.

[9] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, “Predicting elections with Twitter: What 140 characters reveal about political sentiment,” *Proc. Fourth Int. AAAI Conf. Weblogs Soc. Media*, pp. 178–185, 2010.

[10] P. Burnap *et al.*, “Detecting tension in online communities with computational Twitter analysis,” *Technol. Forecast. Soc. Change*, vol. 95, pp. 96–108, 2015.

[11] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, “Twitter power: Tweets as electronic word of mouth,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 11, pp. 2169–2188, 2009.

[12] H. Bakshi, “Framework for Crawling and Local Event Detection Using Twitter Data,” *MSc thesis*, no. May, p. 78, 2011.

[13] A. Bruns, J. Burgess, K. Crawford, and F. Shaw, “qldfloods and @ QPSMedia : Crisis Communication on Twitter in the 2011 South East Queensland Floods,” *Methodology*, no. Cci, pp. 1–57, 2011.

[14] A. Kumar, M. Jiang, and Y. Fang, “Where not to go?: detecting road hazards using twitter,” *Proc. 37th Int. ACM …*, vol. 2609550, pp. 1223–1226, 2014.

[15] V. Lampos and N. Cristianini, “Tracking the flu pandemic by monitoring the social web,” in *2010 2nd International Workshop on Cognitive Information Processing, CIP2010*, 2010, pp. 411–416.

[16] C. Collins, S. Hasan, and S. V Ukkusuri, “A Novel Transit Rider Satisfaction Metric : Rider Sentiments Measured from Online Social Media Data,” *J. Public Transp.*, vol. 16, no. 2, pp. 21–45, 2013.

[17] T. T. B. Luong and D. Houston, “Public opinions of light rail service in Los Angeles , an analysis using Twitter data,” *iConference 2015 Proc.*, pp. 2–5, 2015.

[18] K. Sasaki, “Feasibility Study on Detection of Transportation Information Exploiting Twitter as a Sensor,” *2012 AAAI Work. - Semant. Cities*, pp. 30–35, 2012.

[19] S. Kumar Sharma, X. Hoque, and P. Chandra, “Sentiment Predictions Using Deep Belief Networks Model for Odd-Even Policy in Delhi,” *Int. J. Synth. Emot.*, vol. 7, no. 2, 2017.

[20] S. Si and M. Win, “Target Oriented Tweets Monitoring System during Natural Disasters,” in *Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference*, 2017, pp. 143–148.

[21] A. Giachanou and F. Crestani, “Like it or not: A survey of Twitter sentiment analysis methods,” *ACM Comput. Surv.*, vol. 49, no. 2, p. Article 28; 1-41, 2016.

[22] B. Liu and L. Zhang, “A survey of opinion mining and sentiment analysis,” *Min. Text Data*, pp. 415–463, 2012.

[23] M. Sadegh, R. Ibrahim, and Z. A. Othman, “Opinion Mining and Sentiment Analysis : A Survey,” *Int. J. Comput. Technol.*, vol. 2, no. 3, pp. 171–178, 2012.

[24] V. N. Khuc, C. Shivade, R. Ramnath, and J. Ramanathan, “Towards building large-scale distributed systems for Twitter sentiment analysis,” *Proc. ACM Symp. Appl. Comput.*, pp. 459–464, 2012.

[25] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, “Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph,” *Proc. Conf. Empir. Methods Nat. Lang. Process.*, pp. 53–56, 2011.

[26] F. Å. Nielsen, “A new ANEW: Evaluation of a word list for sentiment analysis in microblogs,” in *CEUR Workshop Proceedings*, 2011, vol. 718, pp. 93–98.

[27] S. M. Mohammad and P. D. Turney, “Crowdsourcing a word-emotion association lexicon,” in *Computational Intelligence*, 2013, vol. 29, no. 3, pp. 436–465.

[28] M. Hu and B. Liu, “Mining and summarizing customer reviews,” *Proc. 2004 ACM SIGKDD Int. Conf. Knowl. Discov. data Min. KDD 04*, vol. 4, p. 168, 2004.

[29] C. D. Manning, J. Bauer, J. Finkel, S. J. Bethard, M. Surdeanu, and D. McClosky, “The Stanford CoreNLP Natural Language Processing Toolkit,” *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. Syst. Demonstr.*, pp. 55–60, 2014.