Cleanliness Campaign V/S Sanitation Related Diseases - Are they parallel in public perspective?

Aarzoo Dhiman
Indian Institute of Technology Roorkee,
Uttarakhand- 247667, India
aarzoodhiman.dcs2017@iitr.ac.in

Durga Toshniwal
Indian Institute of Technology Roorkee,
Uttarakhand- 247667, India
durgafec@iitr.ac.in

Soumya Somani
Symbiosis Institute of Technology,
Pune, Maharashtra- 412115, India
soumya.somani@sitpune.edu.in

Preeti Malik
Indian Institute of Technology Roorkee,
Uttarakhand- 247667, India
parimalik.pcs2016@iitr.ac.in

ABSTRACT

Social media data is playing an important role in healthcare and it is being used for performing many epidemiological tasks such as outbreak surveillance, intervention surveillance, modeling the disease spread through a community, etc. This is due to easy and early availability of social data unlike the clinical data sources which have very limited availability. Twitter data, being in the form of micro-blogs, is the most effective way of performing any study on the thought process of the public as people tweet about anything and everything on their handles. In the present research work, the Twitter data related to an Indian National Level cleanliness campaign, called *Swachh Bharat Abhiyan (SBA)* and the diseases which occur due to lack of cleanliness such as Dengue, Malaria, Diarrhoea, etc. has been collected for the period of 1 January, 2018 to 31 March, 2018. A demographic and temporal analysis of the Twitter data has been performed to compare and contrast the perception of Indian citizens towards SBA and diseases caused by lack of cleanliness. A study of the impact of SBA on occurrence of many diseases which occur due to lack of cleanliness has also been performed. Our experiments showed that the tweets related to both the topics were not very correlated and sentiment analysis of such tweets showed that most of tweets had neutral sentiments.

KEYWORDS

Data Mining, Cleanliness Campaign, Common Water, Sanitation, Swachh Bharat Abhiyan

ACM Reference Format:
Aarzoo Dhiman, Durga Toshniwal, Soumya Somani, and Preeti Malik. 2018. Cleanliness Campaign V/S Sanitation Related Diseases - Are they parallel in public perspective?. In *Proceedings of epiDAMIK*. epiDAMIK, London, UK, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Health related issues can be caused due to many reasons for example, bacterial, viral, fungi, microbial, genetic, parasitic etc. Broadly, the sources of all these reasons can be environmental, social, economic, physical, chemical, political and biological factors etc. Epidemiologists carry out investigations that examine all the above mentioned socio-economic, political and environmental factors that cause health related issues to improve public health. This study is useful in improving the overall public health and health of the disadvantaged, find out the relation between genetic factors, environmental factors and personal behaviors and their interplay, diminish the sources of disease causing agents and study the influence and effects of health programs and services on overall public health.

Multiple air-borne, water-borne and food-borne communicable diseases such as Diarrhoea, Dengue and Typhoid etc. are caused by the lack of proper sanitation, solid and liquid waste management and cleanliness. The occurrence of these diseases can create outbreaks in a few days. That's why, epidemiologists have started working on *Early outbreak detection systems* [2], which are able to detect the outbreaks in the earlier stages of its spread. The primary mode of doing this is by modeling the spread of an outbreak and then predicting the future number of cases by using some available data sources. Traditional data sources have their roots in collecting data through in-patient records of several clinics and hospitals in different regions of the country. However, one primary drawback of these data sources is the delay in availability of data by few weeks or months. Thus, epidemiologists have moved their attention from traditional data sources to web data sources such as social media networks, blogging and micro-blogging networks and search engine query logs. The web data is used because of the ease of access, faster availability and huge amount, which can help in detection of spread of a disease in the earliest stages possible.

Along with monitoring the spread of a disease in a community, epidemiology also deals with providing the measures to eliminate the causes of the spread of the diseases. Such

epiDAMIK, Workshop, London, UK. August 20, 2018

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

measures mostly include the preventive measures for example, vaccination, campaigns and teaching people about risks and abuses of drugs etc. These measures of intervention also need to be monitored to track their effects in the community so that appropriate actions can be taken. One such national level cleanliness campaign, called *Swachh Bharat Abhiyan (SBA)* was launched by government of India on October 02, 2014, to improve the cleanliness situation in India. One primary aim of this campaign is to make India free from open defecation and achieve 100 percent scientific solid waste management by October 2019. However, there are very few statistics provided by the government, which can ensure the level of involvement and awareness among people towards SBA and diseases which occur due to lack of cleanliness. In this paper, first a comparison study of geographic and temporal distribution of the tweets related to SBA and sanitation related diseases has been performed. This study will help in determining the awareness of the citizens of India about the causal relationship between the two topics: SBA and common water and sanitation related diseases.

There are several diseases which are caused due to lack of cleanliness such as Dengue, Malaria etc. However, there is very limited availability of the standard clinical data sources which can provide exact number of cases of these diseases. Hence, Twitter data pertaining to these diseases has been collected and studied to study the impact of SBA on prevalence of diseases caused due to lack of cleanliness for the period of 1, January 2018 to 31, March 2018.

There has been much research work done related to monitoring the outbreak surveillance and tracking the impact of any intervention in a community using the social media data. However, there has not been much research work done which aimed to monitor the relationship between the two. Hence, in this paper, Twitter data is used to monitor the awareness of people on relationship between the cleanliness campaign i.e. Swachh Bharat Abhiyan(SBA) and spread of common water and sanitation related diseases such as Malaria, Dengue etc.

The rest of the paper has been organized as follows. Section 2 contains the related work in the field of epidemiology. Section 3 contains our proposed work which includes the data set description and the methodology used for our study. Section 4 contains results and discussion, which highlights important findings of our analysis. Finally, the paper ends with conclusion and references.

2 RELATED WORK

Social media sites have become the source of providing a variety of features which fulfill many purposes such as social networking, professional networking, media sharing content production, knowledge and information aggregation, virtual reality and gaming environment etc.[9], professional education, organizational promotions, patient care, patient education and spreading information about public health programs. Essentially, Social media allows to ask, and answer, questions were never thought to be possible.

In [1], Rumi Chunara et al. performed early epidemiological assessment using various social media sources during the 2010 Haitian cholera outbreak. Their study showed a good correlation among the official data and social media data which was available up to 2 weeks earlier. Through this study, the authors proposed that social media data can be used in replacement of official data in an outbreak setting to get timely estimates of the disease dynamics. Another approach by J. Gomide et al. [4] studied the extent of Twitter as a tool for surveillance of Dengue epidemic. The methodology proposed in the research work was based on four dimensions: volume, location, time and public perception. First, the public perception dimension was explored by performing sentiment analysis, which filtered out the content that is not relevant for Dengue surveillance. Then, the number of cases reported by official statistics and the number of posts on Twitter during the same time period was correlated and verified. The authors exploited the spatio-temporal dimension of the data to create clusters and the quality of the clusters were then compared to the official data. Another recent study by King-Wa Fu et al. [3] aimed to provide the baseline data model for Zika Virus related English tweets. Its motivation came from the 2015-2016 Zika Virus epidemic in The United States. This study focused on ZIKV-infected pregnancy which could be complicated with fetal microcephaly and long-term developmental disability. As stated in the study, "Epidemiological evidences suggested that ZIKV might cause GuillainBarre syndrome. The World Health Organization (WHO) declared it a Public Health Emergency of International Concern (PHEIC) on February 1, 2016". The authors presented an incidence trend analysis of Zika Virus-related Twitter data and content analysis of a cross-sectional sample of Zika Virus-related English Tweets in their research work.

Now and again, the government keeps on introducing preventive measures to eliminate the socio-economic, environmental, chemical and biological factors behind the causes and spread of a disease to improve public health in a community. The effects of these preventive measures need to be tracked so that appropriate actions could be taken. In 2010, Scanfeld et al.[6] examined the data from Twitter to track the misuse and misunderstanding related to the use of antibiotics in the society by using content analysis techniques. Later in 2017, Shah et al.[7] traced the change in behavior of users search data before and after the introduction of Rota Virus and Noro Virus vaccination in US, UK and Mexico by using the data from Google quantified Internet Query Share (IQS). SBA was first launched in 2014 and since then very less research work has been done related to it and there is no research work done which compares the awareness of these two issues in common public. In 2015, Sahil Raj et al. [5] collected tweets related to SBA and performed simple sentiment analysis to find out perception of Indian citizens towards SBA. Later in 2016, Devendra et al. [8] tested their sentiment analysis tool Senti-Meter on Twitter data related to SBA. They studied 1200 tweets collected for the period of January 2016 to March 2016 and performed manual tagging to evaluate the accuracy of their tool. Both of these works worked on very less number of tweets and did not consider any other demographic details of the Indian cities and states.

3 PROPOSED WORK

The effect of programs like SBA on the occurrence of sanitation related diseases is yet unexplored. To see these changes in the society it's a must that people are aware about the relation between cleanliness and diseases. The primary aim of this study is to track the involvement and perception of people towards SBA and the sanitation and water related diseases. The data related to these topics has been collected separately using the Twitter API and then compared to determine any causal relationships among them.

3.1 Dataset

This section gives some details on the datasets used for the study. The Twitter data related to SBA and the sanitation related diseases has been used to monitor the involvement of people in both these topics and to determine any relationship among them. The datasets have been collected over Twitter Live Stream using Suitable keywords for a period of three months i.e., January 2018 to March 2018.

3.1.1 Disease Data. Tweets for diseases related to common water and sanitation have been collected for a period of three months using Twitter API. Details can be seen in Table 1.

Table 1: Disease Data Description

Sr.No.	Attribute	Values
1	Number of diseases	9
2	Names of diseases	Chikungunya, Cholera, Dengue, Diarrhoea, Hepatitis, Japanese Encephalitis, Malaria, Typhoid, and Zika
3	Tweets Collected	18 thousand

3.1.2 Swachh Bharat Abhiyan Data. SBA related tweets have been accumulated using the keywords elaborated in Table 2. 4 hundred thousand tweets have been collected for the given period with specific numbers given in the Table 2.

Table 2: Description of Twitter data collected using keywords related to SBA

Hashtags	Examples	Tweet example	Number of Tweets
General	Swachh Bharat Abiyan MyCleanIndia	@marineravin: @tavleen_singh anything you wanna more to add abt swachh Bharat Abhiyan ... @sanjayuvacha @amitmehra	322,287
Toilet Related	Open Defecation MyCity-MyPride	@paramiyer_: Congratulations to Team @swachhbharat. Tirunelveli district in Tamil Nadu has been declared #OpenDefecationFree.	26,846
Cities Related	SwachhUP SwachhJhar	RT @lezlietripathy: Participated in Cleaning #Vesave Beach Today. An initiative by @AfrozShah1 Supported by @Dev_Fadnavis @AUThackeray Today. #SwachhBharat #Swachhmaharashtra #swachhversova	24,736
Rural Area Related	ZSBP SbmZSBP	@kishanganjzsbp: Morning follow up and pit digging in Gachpada Panchayat #ZSBP #SwachhBharat #SwachhBihar #SBM-Gramin @SwachhBihar @LSBA_Bihar @swachhbharat	86,868

3.2 Proposed Methodology

Fig. 1 briefly represents the steps of the proposed method. All these steps are explained in the following subsections.

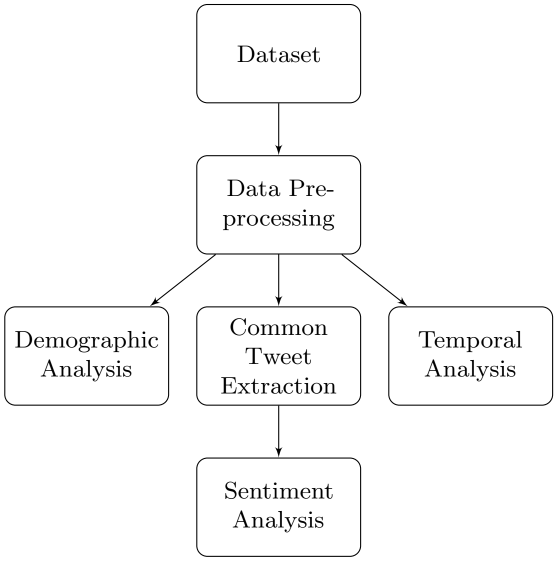

Figure 1: Proposed Methodology

3.2.1 Data Preprocessing. Twitter users need not specify their locations in the account details. This may be the reason why not all the tweets collected have a location attribute in them. Such tweets are needed to be filtered out for further processing so that the locations could be captured.

3.2.2 Demographic Analysis. Let any state or union territory of India be denoted as S_{i}, where

$$i = 1 \text{ to } n,$$

and any disease be denoted as D_{j}, where

$$j = 1 \text{ to } m,$$

Total number of tweets about diseases is denoted by TD,

$$TD = \sum_{i=1}^{n} TD_i$$

where TD_{i} is defined as,

$$TD_i = \sum_{j=1}^{m} TD_{ij}$$

where,

$$TD_{ij} = \text{No. of tweets from state i about disease j}$$

Total number of tweets about SBA is denoted by TS,

$$TS = \sum_{i=1}^{n} TS_i$$

where,

$$TS_i = \text{No. of tweets from state i about SBA}$$

TD_{i} and TS_{i} values are compared to give the state wise distribution.

3.2.3 Temporal Analysis. Let the number of weeks be denoted by k, where $k = 1$ to w. The total number of tweets of any disease be denoted by TD_{j}, where

$$TD_j = \sum_{k=1}^{w} TD_{jk}$$

where,

$$TD_{jk} = \text{No. of tweets about a disease j in week k}$$

The total number of tweets of SBA be denoted by TS, where

$$TS = \sum_{k=1}^{w} TS_k$$

where,

$$TS_k = \text{No. of tweets about SBA in week k}$$

Normalized values of TD_{jk} and TS_{k} values are compared to give a weekly distribution.

3.2.4 Common Tweet Extraction. The datasets are further processed to extract the tweets which talk about SBA and any sanitation related disease at the same time. To extract such tweets we performed a simple keyword search for the mention of both the topics from the Twitter data at the same time. This extraction has been done to study the distribution of the parallel thoughts of people on both the topics.

3.2.5 Sentiment Analysis. Sentiment analysis is a supervised classification process to predict the opinion of a person through the text which is related to some topic. We used sentiment analysis on our Twitter corpus to capture the opinions and sentiments of people towards SBA. Each tweet has been classified into three opinions: positive, negative and neutral by using *Word Sense Disambiguation, Senti Word Net and word occurrence statistics using movie review corpus.* We used the dedicated sentiment classification library of python for our study. If sentiment score value comes out to be greater than 0 then the sentiment is classified as positive, if it comes out to be less than 0 then the sentiment is classified as negative and otherwise neutral. Sentiment Analysis is performed on the tweets which talk about both the topics at the same time to find the opinion of people regarding the two topics.

4 RESULTS AND DISCUSSIONS

In this section, we have highlighted some of the important results of our experimentation. Section 4.1 presents the involvement and emotions about SBA and common water & sanitation related diseases in different states in India. Section 4.2 presents the weekly distribution of tweets for the period of three months. Section 4.3 gives the monthly distribution of common tweets among the SBA data and common water & sanitation related disease data to derive the perception of common public about the relationship among them. All the experimentations have been performed using Python.

4.1 Demographic Analysis

First, we present a state level distribution of total tweets for three months period related to SBA and common water & sanitation related diseases as shown in Figure 2. As visible from the figure, the number of SBA tweets is greater than that of the tweets related to common water & sanitation related diseases. This depicts that the overall SBA related public awareness is higher than that of sanitaion related diseases. It also depicts that in case of SBA, Maharashtra has shown the maximum number of tweets and in case of diseases, Delhi has shown the maximum number of tweets. There are some other states as well where number of SBA related tweets is very high but number of disease related tweets are very less e.g. in case of Madhya Pradesh. This can be due to the fact that Madhya Pradesh has been ranked 1st in SBA rankings 2017 given by government of India.

As seen from Figure 2, number of tweets related to SBA are overshadowing the number of tweets related to the diseases. Hence, we extract sanitation related tweets out of the SBA tweets using the 'toilet' related keywords such as 'open defecation' and 'toilet' etc. Figure 3 gives a state level distribution of sanitation related tweets and sanitation related diseases. The figure depicts that the number of disease related tweets is greater than that of sanitation related SBA tweets. As seen from Figure 2 and Figure 3, the difference is number of tweets in all the three types of tweets is very high. There is very less correlation between the number o ftweets for all the three sets, which means that the people who are talking about one topic may not be talking about the other topic at the same time. This shows that although the overall popularity of SBA is more than the awareness of common water & sanitation related diseases but when it comes to specific reasons behind SBA (i.e. improving cleanliness situation), people are not very aware of its relationship with the effects of SBA (i.e. elimination of causes behind sanitation related diseases).

Further, to make the study exhaustive, Pearson's correlation of the normalized number of tweets (i.e. percentage of number of tweets) related to SBA, sanitation and sanitation and water related disease has been performed. Table 3 gives the correlation value and the P-values for the same. The correlation for most of diseases is found to be negative that means they have an inverse relationship with the number of tweets related to SBA. However, the P-values are mostly greater than 0.05, which may be due to low sample size. This correlation is primarily quantitative in nature. Most of the correlation values are found to be negative. This shows that when there are high number of tweets related to SBA and sanitation, there may be less number of tweets related to some diseases such as Chikungunya, Cholera, Zika etc. The positivity in correlation is also found to be near to zero such as in case of Dengue, Hepatitis and Malaria. The reason behind such uncorrelated behavior may be a few number of

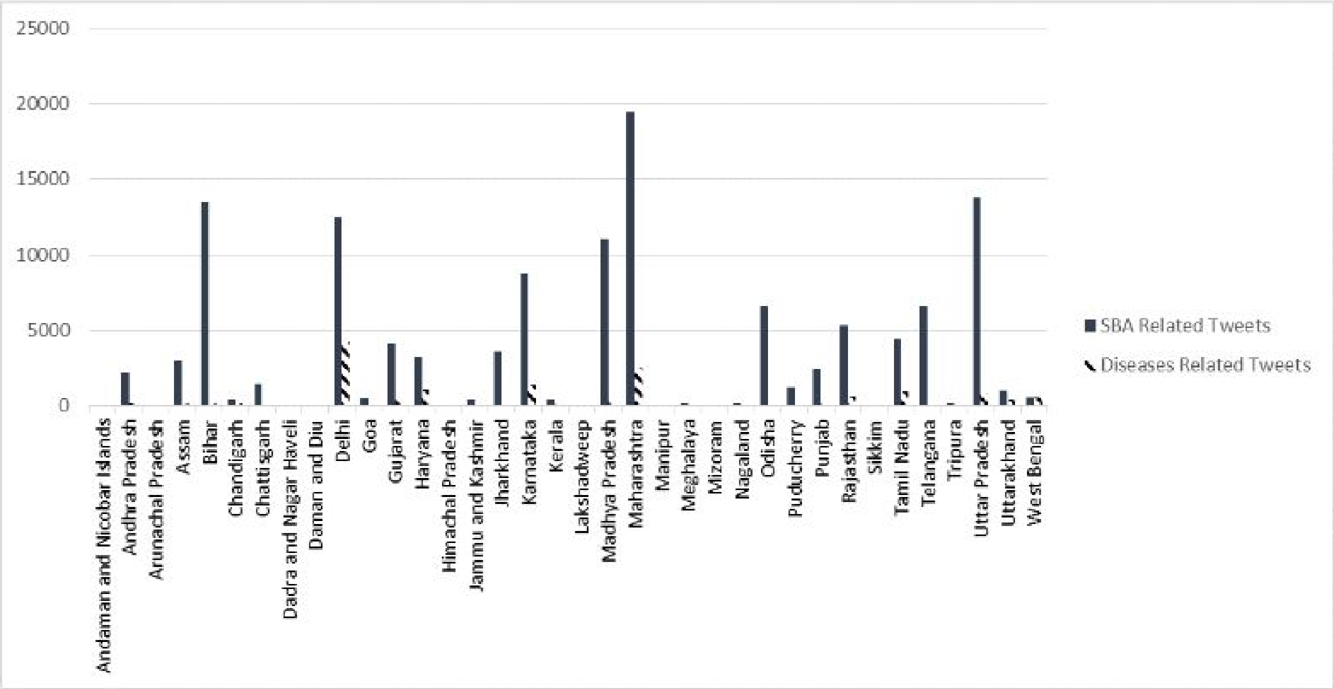

Figure 2: State wise distribution of SBA v/s diseases tweets

tweets but this also signifies that the people do not view both of these topics as correlated.

4.2 Temporal Analysis

The above study gave us the demographic distribution of the tweets related to SBA and sanitation related diseases. Now, to study the change in number of tweets over the given period we have performed weekly trend analysis. Here, the change in percentage of tweets related to these diseases and SBA over the three month period has been analyzed as shown in Figure 4. A large number of tweets related to Dengue, Hepatitis and Malaria are prominent in this duration of the year. Also, the number of tweets related to Dengue and Malaria increases in the month of March which can be due to the increase in mosquitoes in any area. The change in number of tweets related to SBA and sanitation over the three months period can also be seen in the figure.

4.3 Common Tweets Analysis

From the above distribution no real correspondence can be seen between the two topics. To find this, the tweets having both the keywords, from SBA as well as diseases, are found out. As can be seen in Figure 5, the overall number of these tweets which mention SBA and sanitation related disease at the same time are very few. So, people are supporting SBA and talking about health issues individually but there is a lack of awareness among people about how SBA is making a difference in terms of elimination of sanitation and water related diseases. Though the number of tweets are increasing during March because this is the period of occurence of water and sanitation related diseases e.g. Dengue, Diarrhoea, Malaria etc. but these are considerably very low.

4.3.1 Sentiment Analysis. To study the opinions of the people towards both the topics, we extracted the tweets which talk about both the topics i.e. SBA and diseases related to sanitation and performed sentiment analysis on them. There were very few tweets which were talking about both the topics at the same time and the sentiment analysis of these tweets show that most of the tweets show a neutral sentiment. Large number of neutral tweets show that most of the tweets are generally related to spreading awareness about the cleanliness

Table 3: Correlation between the number of tweets related to SBA, sanitation and santitaion and water related diseases

Disease	SBA related tweets	P-value	Santitation related tweets	P-value
Chikungunya	-0.461	0.113226	-0.479	0.097393
Cholera	-0.422	0.150817	-0.516	0.071292
Dengue	0.0175	0.954782	-0.29	0.336281
Diarrhoea	-0.154	0.615615	0.2009	0.510394
Hepatitis	0.0619	0.840839	-0.009	0.977562
Japanese Encephalitis	0.1307	0.670305	-0.046	0.881549
Malaria	0.0109	0.971833	-0.109	0.722413
Typhoid	0.1476	0.630398	-0.026	0.931693
Zika	-0.445	0.127708	-0.579	0.037992

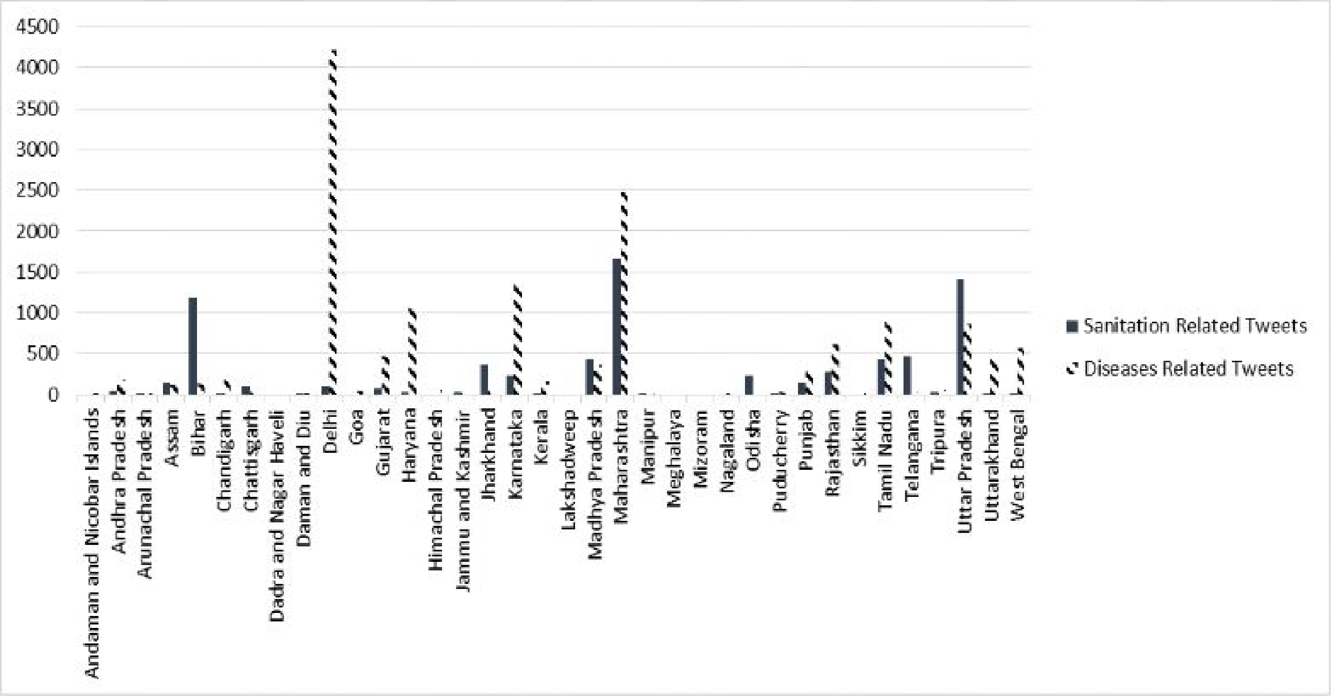

Figure 3: State wise distribution of sanitation related v/s diseases tweets

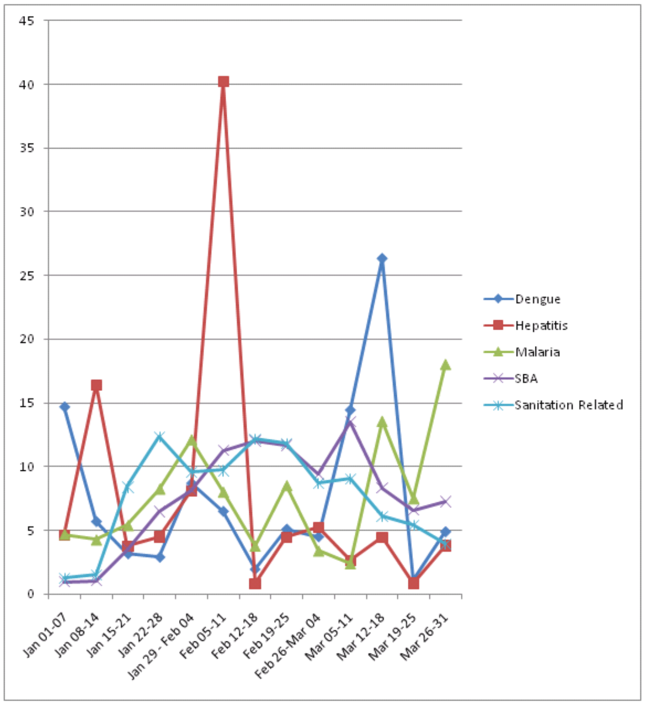

Figure 4: Weekly distribution of percentage of tweets

campaign and its benefits in context to different diseases that are caused due to lack of cleanliness. This supports our previous deduction that people are aware about both the topics separately, but they are not much interested in talking about both the topics as being related to each other.

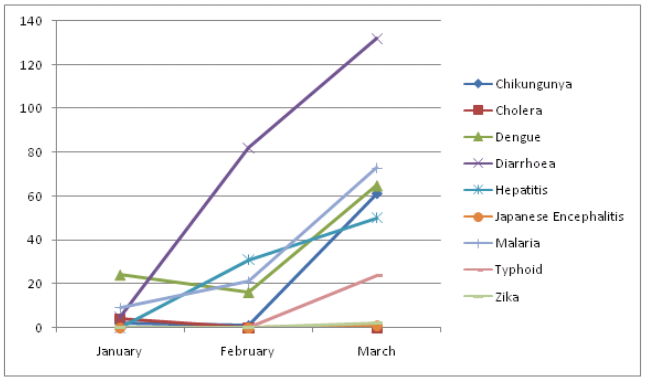

Figure 5: Number of common tweets

5 CONCLUSION

In this paper, the Twitter data is used to capture the insights of public on two topics that have a causal relationship among them i.e. SBA and sanitation related diseases. Through this study, the perception of people about the relationship between these two topics has been monitored. Here, the SBA and disease related data has been analyzed separately as well as collectively. Cleanliness and diseases are well connected terms for real but the results from this work announce otherwise. Results show that people are generally aware of SBA as well as sanitation related diseases on an individual level but they are very less aware about the relationship between the two. This can also be seen as a negative fact as people are tweeting with popular SBA hash-tags without knowing its value and effects in reducing the disease occurrence.

REFERENCES

[1] Rumi Chunara, Jason R Andrews, and John S Brownstein. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American journal of tropical medicine and hygiene* 86, 1 (2012), 39–45.

[2] Ed De Quincey and Patty Kostkova. 2009. Early warning and outbreak detection using social networking websites: The potential of twitter. In *International Conference on Electronic Healthcare*. Springer, 21–24.

[3] King-Wa Fu, Hai Liang, Nitin Saroha, Zion Tsz Ho Tse, Patrick Ip, and Isaac Chun-Hai Fung. 2016. How people react to Zika virus outbreaks on Twitter? A computational content analysis. *American journal of infection control* 44, 12 (2016), 1700–1702.

[4] Janaína Gomide, Adriano Veloso, Wagner Meira Jr, Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. 2011. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In *Proceedings of the 3rd international web science conference*. ACM, 3.

[5] Sahil Raj and Tanveer Kajla. 2015. Sentiment analysis of Swachh Bharat Abhiyan. *International Journal of Business Analytics and Intelligence* 3, 1 (2015), 32.

[6] Daniel Scanfeld, Vanessa Scanfeld, and Elaine L Larson. 2010. Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control* 38, 3 (2010), 182–188.

[7] Minesh P Shah, Benjamin A Lopman, Jacqueline E Tate, John Harris, Marcelino Esparza-Aguilar, Edgar Sanchez-Uribe, Vesta Richardson, Claudia A Steiner, and Umesh D Parashar. 2017. Use of Internet search data to monitor rotavirus vaccine impact in the United States, United Kingdom, and Mexico. *Journal of the Pediatric Infectious Diseases Society* (2017), pix004.

[8] Devendra K Tayal and Sumit K Yadav. 2017. Sentiment analysis on social campaign Swachh Bharat Abhiyan using unigram method. *AI & SOCIETY* 32, 4 (2017), 633–645.

[9] C Lee Ventola. 2014. Social media and health care professionals: benefits, risks, and best practices. *Pharmacy and Therapeutics* 39, 7 (2014), 491.