Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 70 (2015) 801 – 807

Procedia Computer Science

4th International Conference on Eco-friendly Computing and Communication Systems, ICECCS 2015

# An Effective Approach to Track Levels of Influenza-A (H1N1) Pandemic in India Using Twitter

Vinay Kumar Jain[a], Shishir Kumar[b]

[a,b]*Department of Computer Science and Engineering,Jaypee University of Engineering and Technology,Guna,473226,India*

**Abstract**

India is one of the fasted economies in the world, having the second largest population. Communicable diseases in the 21st century remain a major public health threat due to the globalization. The emergence of new diseases and the reappearance of older infectious disease cause great impact towards public health. It is very important to monitor this disease early and take a rapid decision to overcome the damage. In this paper, we suggest an approach based on tweets using Twitter to make a good surveillance system.We examine important issues related to 2015 H1N1 pandemic in India like symptoms, prevention steps and availability of medicine. Our study focused on different parameters to find the relevant information towards Influenza-A (H1N1) and the public general awareness towards it. Our analysis shows the use of twitter may allow the detection of disease outbreaks through analysis of data generated by social media.


Peer-review under responsibility of the Organizing Committee of ICECCS 2015

*Keywords:* Twitter;Swine flu; Influenza; SVM;Naïve Bayes;Random Forest;H1N1

## 1. Introduction

The internet is one of the important resources in the field of surveillance systems for tracking disease outbreaks. It provides an opportunity for low cost time-sensitive sources to be exploited to supplement existing traditional surveillance systems[1].

* Corresponding author. Tel.: +91-7544-267051; fax: +91-7544-267011.
*E-mail address: dr.shishir@yahoo.com*


Peer-review under responsibility of the Organizing Committee of ICECCS 2015
doi:10.1016/j.procs.2015.10.120

In India, surveys are most popular methods in different fields to measure public opinions regarding awareness for diseases.Survey based techniques are costly and time consuming process and are not suitable for infectious epidemic. One of the possible solution to overcome this problem is use of social media which allow rapid exchange of ideas and information using internet-based applications.

Social media has been a primary focus of the information retrieval and text mining field from the start, because it produces massive unstructured textual data and displays user relations in real time. There are a rising number of social media tools and a rapidly growing user base across all demographics. Social networking sites have become fast and low cost communication that enables quick and easy access to public information among potential users. India will account for the third-largest user base on micro-blogging site Twitter at 18.1 million by the end of this year[1].The research firm said growth for Twitter is heavily weighted in emerging markets in India and Indonesia to see the most consistent growth patterns[1].

Delays in identifying the beginning of an infectious epidemic results in a big damage towards a society[2]. Consequently, there is strong interest in reducing these delays. One way to accomplish this is through a good surveillance system, which emphasizes the use of real-time data analysis system which detect and characterize unusual activity for further public health. With the popularity of these social media websites such as Twitter, we are going to monitor the pandemic H1N1 virus during 2015. In this paper, a novel technique based on dynamic keywords from RSS feeds are used to retrieve tweets using Twitter is presented. Sentiment analysis and count based technique is applied on data sets to examine important issues related to Influenza-A (H1N1) pandemic. Our study focused on different parameters to find the relevant information towards a particular disease and the general awareness towards it. Fig. 1 show year wise cases of Influenza-A H1N1 in India. Improving monitoring and surveillance system will definitely overcome big damage towards a society[2] .

Data is collected from Twitter of 60 days started from 1st February 2015 to 31 March 2015. From this data we monitor the important terms use for “H1N1” or “swine flu” over time. Various outcomes are presented using Content analysis of tweets along with examine the most affected states where flu spreading rapidly. Presented technique validate Twitter as a real-time content, sentiment, and public attention trend-tracking tool. An investigation is carried out to those peoples who are aware with swine flu along with people actually being sick in real life. Analysis of data set results in making early warning system by detecting an upcoming spike in an epidemic before the official surveillance systems examined.

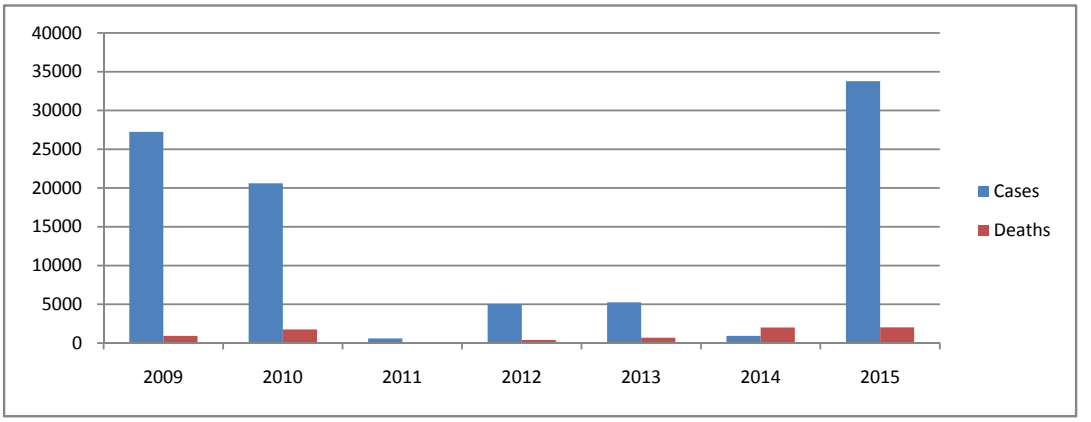


Fig. 1. Year wise cases of Influenza-A H1N1[2]

For India perception we also found a number of advantages that can be found if our monitoring system works well. Estimation of how much resources should be allocated to hospitals and other types of medical services can easily be detected. This could potentially translate into better medical service for the patients, since it's more likely that an adequate number of doctors and nurses are available for a particular region in India where rapidly spreading of the flu taken place. It could also save the medical services money, since it would be possible for the hospitals and

medical services to allocate a suitable amount of personnel on a given day or week. Another advantage with using Twitter for this is that in case it works it would allow us to monitor public health in close to real-time. This can allow us to detect outbreaks of diseases a lot faster than what would be possible to do otherwise.

## 2. Related Work

In the recent years there has been a lot of research into the area of social media data .There are lots of application of social media data in the field of stock market forecasting, election result prediction, movie revenue prediction and customer reviews analysis. There is a limited study towards the public health information system. Some authors worked in the field of public health system using social media data.

Chew and Eysenbach proposed a method during 2009 H1N1 pandemic using Twitter based on specific keywords[3]. Various authors also used some other techniques for finding keywords like Google web search queries related to influenza epidemic[4]. Clustering of Twitter messages by topic and extract meaningful human-readable labels for each cluster is developed by Hu et al[5].Technique based on some specified search terms(flu, vaccine, tamiflu, "h1n1") give high accuracy when applied for in the tweeter related data set.Some authors also used Content analysis and regression models to measure and monitor public concern and levels of disease during the H1N1 pandemic in the United States[5,6].Some other diseases like cholera is also investigated by Chunara et al.[7] to find the cholera outbreak. The authors also developed a framework which provides to quantify users affected by influenza (swine flu) within a community or group and based on user rank algorithm[8].Some authors used machine learning techniques like SVM to predict influenza rates using twitter dataset in Japan[9] .Lamps and Cristianini[6] proposed a method for tracking the various epidemic activity. Some authors also developed tools for real-time analysis of disease using Twitter data showing daily activity of the disease and symptoms[6,11]. Stewart, et al. discussed early warning system with outbreak control and analysis system[11]. Hansen et al. [12]analyzed which tweets attract the biggest attention.

## 3. Public Health Emergencies

A crisis occurred due to epidemic activity caused a massive damage to human life, it is very important for governments and public health agencies to communicate with accurate, timely, direct, and relevant messages to the public by using social media or other broadcasting medium[13]. Diffusion of accurate information during health emergencies is necessary to overcome the impact of epidemic diseases. Social networking sites are one of the easiest and cheaper techniques for spreading information rapidly. This exchange of information regarding immediate health risks can be defined as crisis, risk communication[13].In this paper, we are taking case of Swine flu in India during 2015. It is very important to track the public terminology (Keywords) used during the pandemic as it provides a measure to gain knowledge and public opinions towards the swine flu. An investigation is carried out to know the public opinions and their knowledge related to swine flu. We also analysis how they are expressing themselves, what type of information they are spreading. We also analysis the services and facilities like medicine, testing labs, paramedical staff provided by the government towards the effective ones.

## 4. Data and Methodology

The proposed method is based on the analysis of tweets and web search queries.Started by identifying tweets and queries that are relevant for indicating the presence of flu or flu symptoms. For a good query we have some important and relevant keywords. Data collection methodology is different from other authors[6,15]; we considered those keywords which are popular in news articles. For getting keywords from the news we analysis top newspapers RSS and find out the mostly occurred keywords using N-gram based techniques. This methodology gives dynamic keywords which are popular during a particular time period and related to public sentiments.

### *4.1. Identification of relevant keywords*

To identify those keywords that are relevant in the context of assessing the sentiment of sentences we used two method (a) Keywords from tweets (b) Keywords from RSS feeds.

(*a) Keywords from tweets:* TF-IDF[14] based approach is used to measure the importance of words (or "terms") in a document based on how frequently they appear across multiple documents. Weight is composed by two terms: the first computes the normalized Term Frequency (TF) and the second is to find Inverse Document Frequency (IDF) given in Eq. (1) and Eq. (2).

$$\text{TF (t)} = \frac{\text{(Number of times term t' appears in a document)}}{\text{(Total number of terms in the document)}} \quad (1)$$

$$\text{IDF(t)} = \log\_e \left( \frac{\text{Total number of documents}}{\text{Number of documents with term t in it}} \right) \quad (2)$$

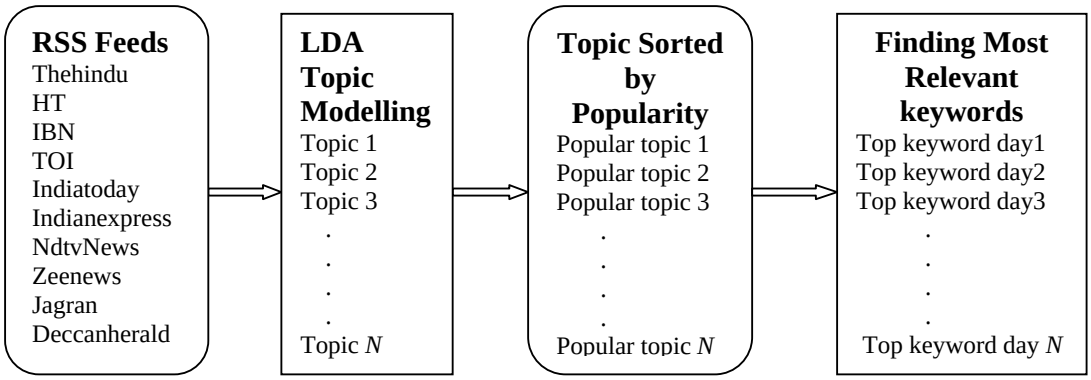


Fig.2. Finding keywords from RSS feed

*(b) Keywords from RSS feeds*: RSS feeds of top newspaper agency are collected and filtered out the main headlines for H1N1 or Swine flu. LDA[15] technique is used to divide news headline to find most popular news for a particular day. From these popular topics we find the top keywords of the day.Fig.2 show the systematically diagram for finding the popular keywords.

Table 1. No. of tweets collected weekly (Feb 2015 to Mar 2015)

| Weeks | No. of tweets |
|---|---|
| 01 Feb-07 Feb 2015 | 7020 |
| 08 Feb-14 Feb 2015 | 10504 |
| 15 Feb-21 Feb 2015 | 22089 |
| 22 Feb-28 Feb 2015 | 17653 |
| 1 Mar-07 Mar 2015 | 10674 |
| 08 Mar-14 Mar 2015 | 7542 |
| 15 Mar-21 Mar 2015 | 8530 |
| 21 Mar 28 Mar 2015 | 7483 |

We collected the tweets dynamically from 01 February 2015 to 28 March 2015 using relevant keywords. A total of 91495 tweets is collected. Table 1 shows no. of tweets collected weekly during the time period. From Fig 3 we can analysis that during 15 February 2015 to 28 February 2015 tweets related to swine flu increases.This shows that the public is more concerned during this time span. After this time period a decline regarding tweets related to swine flu.

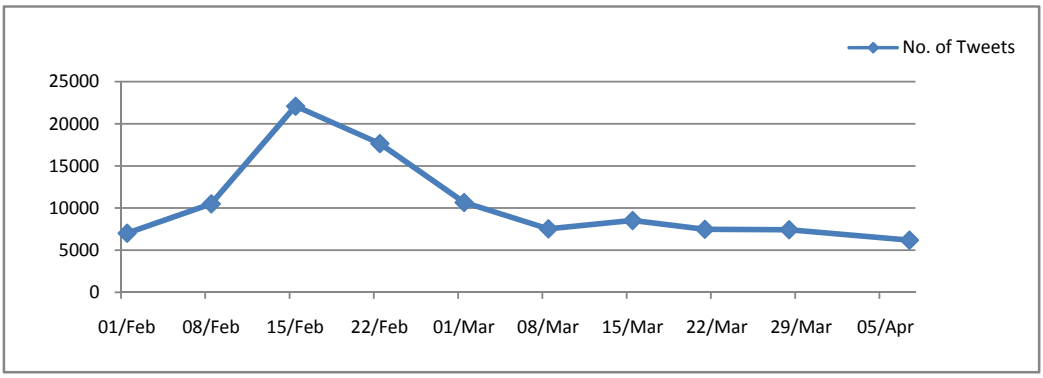


Fig. 3. No. of tweets collected weekly (Feb 2015 to Mar 2015)

### *4.2. Analysis Top Keywords*

In this section we used the count based technique to know the no. of occurrence of a given word in our data set. After analysis, we found that people use various vocabularies for Influenza-A (H1N1), they used swine flu mostly as compared to H1N1 or influenza. Table 2 represents 20 most popular keywords.There are lots of other keywords related to symptoms, prevention and medicine for swine flu also occurred.

**Symptoms**= {cold symptom, respiratory failure, cough, asthma, runny nose, blocked nose, problem in of breath, breathing difficulties, breathing trouble, pneumonia, sore throat, bronchitis, pain in the chest, tonsillitis, vomiting, abdominal pain, dizziness}
**Prevention**= {Tulsi, Kapoor, Mask, face mask, Wash your hands, Avoid touching your eyes, nose, and mouth}
**Medicine**= {Tamiflu, flu vaccine, garlic, ayurvedic medicine, homeopathy medicine, turmeric, Tulsi, Neem}

Table 2. Top 20 Most Frequently Occurring Words

| Word | Occurrence | Word | Occurrence |
|---|---|---|---|
| SwineFlu | 12015 | Sick | 1063 |
| flu | 7471 | Health | 1053 |
| swineflu | 6337 | Mumbai | 804 |
| Swine | 5318 | Outbreak | 692 |
| Flu | 3365 | influenza | 671 |
| swine | 3294 | Rajasthan | 624 |
| virus | 2299 | Kashmir | 604 |
| deaths | 1975 | feel | 553 |
| Swineflu | 1788 | Andra | 425 |
| Gujarat | 1117 | Maharashtra | 388 |

### *4.3. Tracking Public Interest with Twitter Data*

We analysis data top keywords like h1n1, swine, flu or influenza from our twitter data set during different time intervals given in Fig. 4. These swine flu related tweets represent at most just over 15% of the sample tweet volume, and this percentage declined rapidly over time, even as the number of reported H1N1 cases continued to rise.

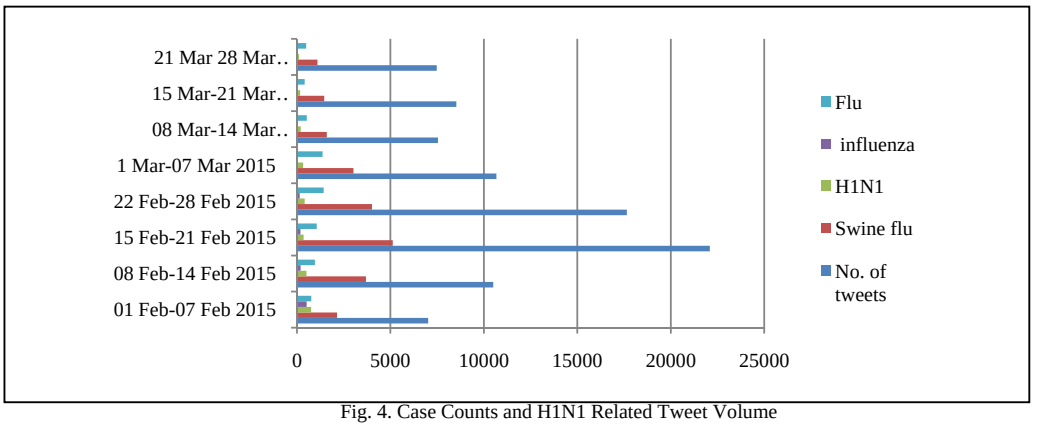


Fig. 4. Case Counts and H1N1 Related Tweet Volume

### *4.4. Classification of Tweets*

Pre-processing the tweets are carried out such as stop word removal, stemming etc. We applied classification to differentiate real-time data from noise or irrelevant tweets. Thus, the purpose of this step is to decrease the amount of noise from the tweets and filter out as irrelevant tweets. Every relevant word related to swine flu is considered as features and a various classification techniques such as SVM, Naïve Bayes, Random Forest and Decision Tree are applied. We tested in to evaluate which technique would produce better results.

Table 3. Comparison of classification techniques

| Classifier | F-Measure | Precision | Recall |
|---|---|---|---|
| Naïve Bayes | 0.77 | 0.70 | 0.86 |
| SVM | 0.77 | 0.70 | 0.84 |
| Random Forest | 0.70 | 0.66 | 0.75 |
| Decision Tree | 0.70 | 0.67 | 0.74 |

The performance analysis of the different classifiers was compared using the dataset of 21040 tweets, covering the period from 15 February 2015 to 15 March 2015. An SVM classifier gives better results as compared to other classifiers with an F-measure of 0.72 is given in Table 3. After feature selection, the best results were obtained for a set of 200 features, achieving an F-measure of 0.77 with both SVM and Naïve Bayes classifiers.

## 5. Conclusion

Twitter offers unique challenges and opportunities for monitoring and surveillance public health. We have presented a method for tracking the flu epidemic in India during 01 February 2015 to 01March 2015 using the contents of Twitter. Our approach could give rapid information in various situations like symptoms corresponding swine flu, prevention techniques used by the user and awareness about the medicine and labs, but mostly can give timely information to government and health agencies. In this study, we find that public opinion towards H1N1 flu

along with a monitoring scheme based on keywords from RSS feeds to analysis more tweets and classify them as relevant or irrelevant. Our results support that the use of social media for tracking a disease will be utilized for knowing the public health in the society.

## References


1. Emarketer. http://www.emarketer.com;2015.
2. Preventive Measures for Swine Flu.http:// pib.nic.in/newsite/PrintRelease.aspx?relid=115710;2015.
3. Chew C M. *Pandemics in the age of twitter: A content analysis of the 2009 h1n1 outbreak*. Master's thesis, University of Toronto;2010.
4. Signorini A . *Social web information monitoring for health* ; 2009.
5. Hu X.,Lei Tang, and Huan Liu. Enhancing accessibility of microblogging messages using semantic knowledge. *In Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, New York, USA, ACM; 2011,p. 2465-2468.
6. Vasileios Lampos and Nello Cristianini.Tracking the flu pandemic by monitoring the social web. *In 2nd IAPR Workshop on Cognitive Information Processing (CIP 2010)*,IEEE Press,;2010, p. 411–416.
7. Chunara R, Andrews JR, Brownstein J S. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *Am J Trop Med Hyg* ;86,(1),2012 , p.39-45.
8. Tang X and Yang C C .Identifying Influential Users in an Online Healthcare Social Network, *IEEE, ISI 2010*,Vancouver, BC, Canada; 2010,p. 43 - 48
9. Aramaki E, Maskawa S, Morita M .Twitter catches the u: detecting inuenza epidemics using Twitter.*In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*;2011,p. 1568-1576.
10. Lee C, Kwak H, Park H, and Moon S. Finding influentials based on the temporal order of information adoption in twitter. *In Proceedings of the 19th International Conference on World Wide Web (WWW'10)*; 2010, p.591–600.
11. Stewart A and Diaz E. Epidemic intelligence: for the crowd, by the crowd .*In Proceedings of the 12th international conference on Web Engineering, ICWE'12*,Berlin, Heidelberg,Springer-Verlag;2012, p.504–505.
12. Hansen L K ,Arvidsson A, Nielsen F A, Colleoni E and Etter M. Good friends, bad news - affect and virality in twitter.Future Information Technology, Communications in Computer and Information Science;2011, p. 34-43.
13. Glik D.*Risk communication for public health emergencies*. Annual Reviews of Public Health;2007,p.33-54.
14. TFIDF.http://www.tfidf.com .
15. Blei D M ,Ng A Y,Jordan M I. Latent Dirichlet Allocation ,*Journal of Machine Learning Research;* 3,2003,p. 993-1022.