ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

Prediction Model for Influenza Epidemic Based on Twitter Data

Sangeeta Grover[1], Gagangeet Singh Aujla[2]

Department of Computer Science and Engineering, Chandigarh Engineering College, Landran, Mohali, India[1]

Department of Computer Science and Engineering, Chandigarh Engineering College, Landran, Mohali, India[2]

Abstract: Today controlling the influenza outbreak has become an important issue of health authorities worldwide to get rid of the epidemic as early as possible. In this research work we have done an associated study of algorithms and methods, modelling the outbreak of any epidemic with the focus of swine flu which must be prevented at an early stage if spread. In the Introduction section we have given the significance of the study with respect to micro-blogging websites like Twitter, Facebook, etc. that studies Social media platforms. In Related Work, we have done a survey from different resources and ideas applied to predict and detect the outbreak of epidemics and studied their advantages and limitations. After that new idea is proposed which can overcome the limitations of models have been proposed. This proposed model comprises a machine learning technique in order to make a model trained and we proposed a new idea of Swine Epidemic Hint algorithm which will look after epidemic activities happening on the Twitter and the Markov Chain state model to categorize epidemic activities into three stages (Beginning of Epidemic, Spread of Epidemic, Decay of Epidemic). Finally, we have proposed a new framework to model epidemic prediction based on the scope of improvements of previous work done.

Keywords: Twitter APIs, Markov Chain State Model, BOWs, Time series classification.

I. INTRODUCTION

Influenza is an infectious disease caused by the influenza virus that is a rapid viral infection that spreads easily one to another person by coughing, sneezing etc. It is a type of variant virus which causes of 250000 to 500000 deaths worldwide every year [1] [4] [5] [6]. It spreads based on multiple vectors like demographic, response to disease, etc. There are many disasters occurred in the past are live examples show the result of an epidemic, for example, in 1918 an epidemic named "Spanish Flu" caused the death of 50-100 million people. Decreasing the effect of seasonal epidemics such as the H1N1 is the main anxiety of the health authorities. Although Researchers have made many medicines to control this, but medicines just ease the symptoms not cure the problem generally because it spreads very rapidly and very difficult to get rid of. Vaccination is said to be the most effective way to prevent the infection of the flu [3] only if, the problem is detected early.

The Social media plays a better role to inhibit and mitigate the hold of influenza like H1N1 as compared to the traditional media because the social media gives the real time information, but the traditional media gives information when anything big event occurs that is not a defensive measure because in this case even a small delay is dangerous, for example, if we depend only on the traditional media, the disease would spread on a very faster rate just because of not detecting the epidemic at the right time that is very arduous to cure. Early discovery is only possible using social media, i.e. micro-blogging websites like Twitter, Facebook, MySpace. As social media is performing an essential role in our day to day life. These social media platforms have the most prevalent channel for a person to express his opinion and feelings by which he/she interacts with millions of social users around the world, [9] for example, health conscious people share their diet plans on social media, not only the diet plans like "How much calories we burnt today" or "how to stay healthy", they also share when and how they fall sick due to some 'xyz' reason. Everyone is sharing a lot of information related to politics, culture, news and disease spread. In the other way, it can be said that social media provides a virtual network that allow people to interact with each other via the Internet. All information (status, tweets, etc.) stored in social communities can help to discover the data related to the epidemics. This information includes healthy (accurate) and false (inaccurate, inconsistent or insufficient) information. It is required to ensure that the information given by the user should be accurate and can be used to stimulate positive health information and health outcome of people is improved rather than false information that may threaten the public safety [9]. Since information is power and strong preventive measures can be taken before the disease spread to the parts of the infected region, either in the form of suggested measures using social communities or by operational health services.

Twitter has become a popular medium for people to share their status (tweets) according to their mood, health issues, relationships, etc. these tweets are also updated by mobile phones by which we can trace the user's exact location and weather condition. Twitter provides the free APIs from which a sampled view can easily be obtained. Twitter APIs allow us to access only 1% sample of the Twitter data which is downloaded depends on the sampling technique we use. Twitter helps to build a map spread model of disease. There are approximately 10 million active twitter users in India who frequently tweets via laptops, mobiles, iPods, etc. While collecting the data

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

twitter gives untried information and data source by which we can extract the onset of flu epidemics and its spread [1]. Special attention is given to the users tweet their post like "High Fever", "I got FLU", "Swine Flu", "H1N1" etc. The accuracy level of the model is evaluated by comparing the model's results to CDC (Centers for Disease Control and prevention) data which is the actual data, obtained by cases of influenza registered manually.
In the following section we are discussing the related work of the prediction model.

II. RELATED WORK

In 2013, **H. Jiangmiao et al. [1]** provided a model to detect flu transmissions. The model is based on "Sina Weibo", a Chinese micro-blogging website which is like a hybrid of twitter and Facebook. They collected over 35.3 million tweets shared by all metropolitan cities in China. Data was extracted on the basis of filtering techniques based on CDC ILI definition. They collected information related to infection centre, the city set, target city, related city and co-related city by using Dynamic Bayesian Network. They depicted their results of detection and transmission at city level.
In 2012, **L. Bumsuk et al. [3]** looked on tweets from twitter and compare the tweet corpus to Influenza- like Illness (ILI) data sets, weather factors, and the flu forecast. Comparison was made on the daily four-level flu forecast from the Korea Meteorological Administration (KMA) and the weekly ILI proportion Korea Centre of Disease Prevention (KDLC). In their results they depicted comparison graphs of flu signals on twitter to weather factors and flu forecast.
In 2011-2012, **A. Harshvardhan et al. [4] [5] [6]** Had proposed a SNEFT architecture model which contained crawler, predictor and detector components to predict influenza activities using information gathered from micro-blogging websites like twitter and Facebook. In this framework, Auto Regressive Moving Average (ARMA) model was used to predict ILI incidences. This model worked with certain accuracy. Tools mainly used in this model were ARMA model, ARX model, OSN crawler. Tweets were collected from date 18th Oct 2009 to 31st Oct 2010 and recorded 4.7 million tweets from 1.5 million unique users along their social relationships from twitter. Authors provided Hourly and weekly basis results by using this model. Similarly, **C. Aron [7]** analysed 500 million messages from twitter which took 8 months time and used filtering and regression. They obtained 95% correlation between their results and national health statistics.
S. Takeshi et al. [8], the authors created an earthquake reporting framework to detect earthquake activities. That framework explored the real-time nature of Twitter, specifically for event recognition. Semantic analyses were used in tweets to characterize them in positive (tweets related to the occurrence of the earthquake) and negative classes. The SVM (Support Vector Machine) which is a machine learning algorithm was used to train the database by giving positive and negative examples to the machine. Author proposed a model which considered each Twitter client as a sensor, and on those sensory observations, earthquake events were detected. To detect relevant activities Area estimation strategies, such as Kalman filtering and particle filtering were utilized to quantify the areas of occurrence of the event.

T. Xuning et al. [9] had proposed a framework that was used to quantify users affected by influenza (swine flu) within a social networking community and introduced an UserRank algorithm that incorporated the link structure, content similarity, responding order and time of repliers. They tested for swine flu forums which were of small size had 12 authorized members and each of them had 15.6 friends on average. There were 90 threads in total they tested. They give their results with 100% precision.
L. Vasileios et al. [10], Authors proposed a method for tracking the epidemic activity. Tweets on Twitter in the UK over 5.5 million of users were observed in order to extract the calculations from twitter that measured the diffusion of ILI among the various regions and tested the activities for 24 weeks and depicted 95% accuracy when compared to the official health reports of HPA (Health Protection Agency) and the correlation coefficient's table of twitter's flu score and HPA's score and correlation graph comparing both.
Recently, an abundance of researchers have been working on the detection of the epidemics using social communities like Twitter, Facebook, etc., to collect real time data that can help us to prevent and detect epidemics. Besides, many mathematical models have been made, but still there are some limitations, for example, lack of real time information, machine learning tools, that can be used for predicting and detecting the epidemic.

III. RESEARCH GAP

Till now, the work related to the detection of the influenza epidemic based upon social networking communities which has been done on a very large scale and different models also have been introduced. However, there is ample scope of improvement of these methods due to the inherent nature of these methods, accuracies and applicability. Hence, the following gap has been found as limited work has done in the following area:

- A learning system should be trained in order to automatically discover keywords which are more useful to predict the ground truth rate.
- Artificial intelligence can be used in combination with probabilistic models like Markov Chains instead of building influenza term corpus only so that model accuracy will be improved.
- A model can be made which will work on a variety of epidemics (if their symptoms are different) instead of one disease detection.
- Work can be done in order to make model language independent.
- A model can be made which can detect and predict epidemic by considering many social media platforms.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

IV. PROPOSED WORK

After conducting systematic literature survey and studying material associated with this problem area, we propose a new framework in which our model comprises different techniques like machine learning algorithms and time series classifications and prediction. This work can be used to overcome the limitations of the previous work has been done. In this model we will build a time series classifications and prediction for identification of the epidemic stage based probabilistic model of vocabulary (BOWs) used in different stages of the epidemic shared online by the act of tweeting.

V. IMPLEMENTATION

In this section we depicted the working of proposed work. The implementation of the proposed framework is described in the following steps:

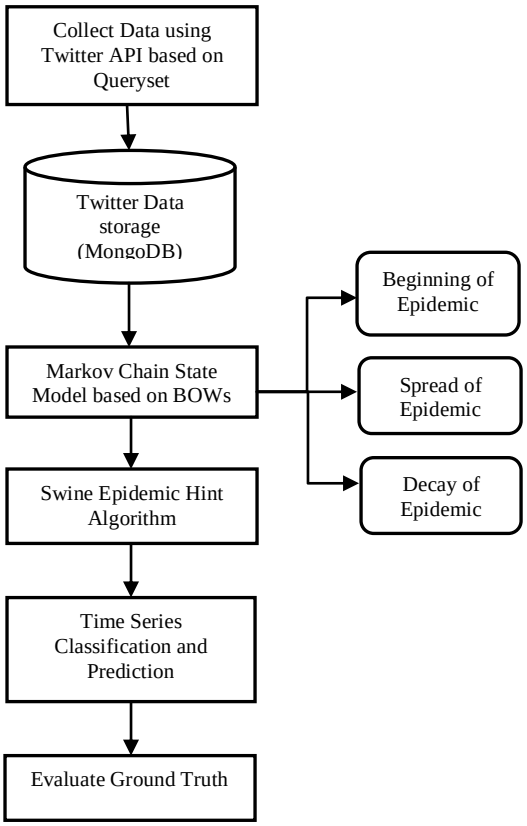

Figure 1: Proposed model and its components

A. Collect Data from Twitter API based on Query Set

Typically, when user tweets, the words used reflect the stage of epidemic spread, these are based on the word/phrase/sentence relevancy score and the stage of epidemic factors that may be accounted for the logic fro detection of the epidemic start. Hence, the query set, which helps to extract relevant tweets and numerical factors helps to calculate the relevance of text are considered as a feature set for doing the research work.

Twitter's prevalence as a data source has prompted the advancement of applications and research in different areas. As Twitter is utilized to stay connected us with our followers and users, we follow and share tweets with each other through the world, and it is also used to give situational attention to an emergency circumstance. Many researchers have utilized Twitter to predict events like earthquakes and distinguish relevant clients to look after to get calamity related data.

A sample of Twitter data can easily be obtained through the APIs which is freely available, to obtain the full view is difficult because the Twitter APIs only allow us to access 1% sample of the Twitter data, which is the result of sampling strategy we will use related to our required information. While collecting Twitter data (runtime) a Query set will be applied, which specifies a set of keywords related to Swine Influenza Activities, so that only useful information stores in MongoDB.

B. Twitter Data Storage (MongoDB)

After collecting the information (tweets) we have to store it somewhere for which we can choose any database, for example, MS access, MySql, Oracle and MongoDB. In this model we choose MongoDB Database because of its performance and fulfilment to the following principles:

- Document-Oriented Storage- MongoDB stores its data in JSON-style objects, which makes it very easy to store raw documents from Twitter's APIs.
- Index Support- MongoDB makes it easy to create indexes optimized for your application because it allows for indexes on any field
- Straightforward Queries- MongoDB's queries, while syntactically much different from SQL, are semantically very similar. In addition, MongoDB supports MapReduce, which allows for easy lookups in the data.

C. Markov Chain State Model based on BOWs

Once, the collection of Tweets database set is ample with a simple size six months of data, the next step is to build the 'Bag Of Words' for each Markov State(Beginning, Spread and Die) through which an epidemic undergoes. So Markov state model can be applied on data with respect to BOWs (Bag Of Words), these words, sentences, phrases, verb, adverb, noun verb, pair combinations show the state of affairs in terms of vocabulary tweeted by Twitter handler to its network which will extract the useful content and will divide it into three states as given below-

- **Beginning of Epidemic**, this state will indicate the beginning stage of the epidemic.
- **Spread of Epidemic**, this state will narrate that the epidemic is spreading or has already spread in the specific area.
- **Decay of Epidemic**, this state indicates that the epidemic in now under control or died.

D. Swine Epidemic Hint Algorithm

In this algorithm each relevant tweet text will be tokenized, and then slop words will be removed and last but not the least stemming will also be done. Once this step is complete, the numerical analysis of the tweet will start based on the following numerical formulas, which will check the relevance of each tweet with respect to the state (Beginning of epidemic, spread of epidemic and decay of the epidemic) tweet will be having. As per above flow chart, the following table describes how the Swine

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

Epidemic Hint Algorithm (SEHA) will calculate the hint score:

TABLE 1
Epidemic Hint Algorithm description

S.no.	Name	Description
A.	BBOW(Beginning Bag Of Words)	Indicates Beginning Stage of epidemics.
B.	SBOW(Spreading Bag Of Words)	Contains the keywords related to the spread of Epidemic.
C.	DBOW(Decay Bag Of Words)	Shows the decay of Epidemic.
D.	Total Bag Of Words	A+B+C
E.	Like Score (Greedy Score)	Total Token found relevant to A, B and C separately using like operator and greedy approach.
F.	Equal Score (Exact Score)	Total Token found relevant to A, B and C separately using equal operator (Exact score).
G.	Precision	F/E
H.	Recall	F/(A+B+C)
I.	Swine Hint Score	(G*H)/(G+H)

BOWs are the simplified representation used for information retrieval [14]. The combination of words/sentences/keywords/phrases represents a Bag (multiset). The aim of Bag Of Words is to retrieve those tweets containing keywords related to Swine epidemic activities with ease, speed and accuracy. We have three BOWs, each one has its own significance, i.e. BBOW(Beginning Bag Of Words), has the keywords that indicate the beginning of the epidemic, SBOW(Spread Bag Of Words), keywords stored in it give hint that epidemic is spreading or has already spread and DBOW(Decay Bag Of Words), comprises information concerning to the decay of the epidemic.

The Like Score is the Greedy Score in which "like operator" is used for calculating term frequency with respect to the bag of words for each Markov states of the epidemic . This operation basically counts the words which are somewhat like the words/phrases in BOW sets. However, in case of the Equal score, exact sentence/word must match with BOW sets to find term frequency.

Tweets are divided into two categories one is "accurate" which is relevant tweet and second is "not accurate" which is not relevant. Precision is defined as the relevant tweets retrieved by equal operator divided by the total number of tweets retrieved by the search operator (like). Recall is defined as the relevant tweets retrieved by Equal operator divided by the total number of keywords comprised in the BOWs.

The Swine hint score is calculated at the end which is the final score of Swine Epidemic Hint Algorithm stated mathematically as the product of precision and recall is divided by the sum of precision and recall i.e.

(Precision * Recall) / (Precision + Recall)

E. Time Series Classification and Prediction

In this research, we may basically work on the concept of Motif Recovery Algorithm for the time series based classification of all the Markov states of the epidemic. While using the approach, machine learning algorithm is also incorporated and machine learning algorithm benefits can be reaped; since our data (Tweet text) is time dependent.

However, it must be done by removing the temporal ordering of individual inputs. Once the data is transformed, multi-linear regression algorithm may be applied.

F. Evaluate Ground truth

After all phases of detecting and predicting the Swine epidemic processed, the model will proceed for evaluation and ground truth validation using Delphi technique [15].This would finally build Inter rater agreements based on the scores given by the Swine Epidemic Hint Algorithm automatically.

VI. CONCLUSION

After doing the deep study of epidemic models and methodologies, we proposed a new model which has not been used yet and will overcome the drawbacks of previous work has been done before. This model will include machine learning algorithm in which system will be trained in order to capture and react to influenza like activities so that the preventive measures can be taken to get rid of the epidemic as early as possible. In this model we will use Twitter APIs to collect the tweets from the Twitter based on the Query set. These extracted tweets are going to store in MongoDB (like the Big Data concept is there). Tweets collected from Twitter will be further categorized in three Markov states (Beginning of the epidemic, Spread of the epidemic and decay of the epidemic). On these Markov States, a Swine Epidemic Hint algorithm will be applied for calculating the score of the tweet. At the end Ground Truth will be evaluated using the Delphi technique which would finally build an Inter rater agreement based on the scores given by the Swine Epidemic Hint Algorithm automatically. As per our estimation this model will give better accuracy and will help in the prediction, detection and control the spread of disasters that come in the future.

REFERENCES

[1] J. Haung, H. Zhao and J. Zhang,"Detection Flu Transmission by Social sensor in China," IEEE International Conference on Green Computing and Communication and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, 2013.

[2] Z. Yi, J. Yueming and L. Wanquan, "Chaos control for a class of SIR epidemic Model with seasonal fluctuation," Proceedings of the 32nd Chinese Control Conference July 26-28, 2013.

[3] B. Lee, J. Yoon, S. Kim and B. Hwang, "Detecting Social Signals of Flu Symptoms," 8th International Conference Conference on Collaborative Computing: Networking, Applications and Worksharing , Collaboratecom,2012.
[4] H. Achrekar, A. Gandhi, R. Lazarus, S. Yu and B. Liu, "Predicting Flu Trends using Twitter Data," The First International Workshop on Cyber-Physical Networking Systems , 2011.
[5] L. Chen, H. Achrekar, B. Liu and R. Lazarus, "Introducing SNEFT – Social Network Enabled Flu Trends," ACM Workshop on mobile Cloud Computing & Services: Social Networks and Beyond (MCS), San Francisco USA, 2010.
[6] H. Achrekar, A. Gandhi, R. Lazarus, S. Yu and B. Liu, "TWITTER IMPROVES SEASONAL INFLUENZA PREDICTION," In HEALTHINF, 2012.
[7] A. Culotta, "Detecting Influenza outbreaks by analyzing Twitter messages," (Submitted on 24 Jul 2010 Department of Computer Science, Southeastern Louisiana University Hammond, LA 7402).
[8] T. Sakaki, M. Okazaki and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors,"WWW2010, April 26-30, Raleigh, North Carolina,2010.
[9] X. Tang and C. C. Yang, "Identifying Influential Users in an Online Healthcare Social Network," IEEE, ISI 2010, May 23-26, Vancouver, BC, Canada, 2010.
[10] V. Lampos and N. Cristianini, "Tracking the flu pandemic by monitoring the social web," 2nd International Workshop on Cognitive Information Processing, 2010.
[11] H. Becker, M. Naaman and L. Gravano,"Beyond Trending Topics: Real-world Event Identification on Twitter," Association for the Advancement of Artificial Intelligence (www.aaai.org), 2011.
[12] M. Naaman, H. Becker and L. Gravano, "HIP AND TRENDY: CHARACTERIZING EMERGING TRENDS ON TWITTER," JASIST, the Journal of the American Society for Information Science and Technology,2011.
[13] S. Kumar, F. Morstatter, & H. Liu, "Twitter Data Analytics," Springer New York, 2014.
[14] Sivic, Josef, and A. Zisserman. "Efficient visual search of videos cast as text retrieval," Pattern Analysis and Machine Intelligence, IEEE Transactions on31.4 (2009): 591-606.
[15] Rowe, Gene, and G. Wright. "The Delphi technique as a forecasting tool: issues and analysis," International journal of forecasting 15.4 (1999): 353-375.

BIOGRAPHIES

Sangeeta Grover (20-05-1989), Sirsa, Haryana, India. He received the B.Tech degree in Computer Science and Engineering from Maharishi Markandeshwar University, Mullana, Ambala, India, in 2011. She is currently an M.Tech student in the Department of Computer Science and Engineering in Chandigarh Engineering College, Landran, Mohali, Punjab, India. Her research interests includes Data mining and Machine learning.

Gagangeet Singh Aujla (15-07-1982), Kapurthala, Punjab, India. He received the B.Tech degree in Computer Science and Engineering from Punjab technical University, Jalandhar, Punjab, India, in 2003, and the M.Tech degree from Punjab Technical University, Jalandhar, Punjab, India, in 2012. He is currently an Associate Professor in Department of Computer Science and Engineering at Chandigarh Engineering College, Landran, Mohali, Punjab, India. His current research interests include Vehicular Adhoc Networks, Wireless Body Area Networks, Cryptography and Wireless Communication. He is a lifetime member of Indian Society of Technical Education and a senior member of Computer Society of India. He currently serves as the reviewer in International Journal of Security and Network Communication, Wiley Publications.