International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-7,October 2015
ISSN: 2395-3470
www.ijseas.com

PREDICTION MODEL FOR INFLUENZA KNN ALGORITHM BASED CLASSIFICATION ON TWITTER DATA

[1] **D.Kavitha** MCA
[1] M.Phil Research Scholar,
[1] Dept of Computer Science,

[2]**Dr.P. Senthil Vadivu**,
[2] Associate Professor/ Head,
[2] Dept of Computer Applications.

[1&2] Hindustan College of Arts and science, Coimbatore.

Abstract— *Today dominantly, the grippe eruption has become a crucial issue of health authorities worldwide to induce elimination of the epidemics as early as attainable. During this analysis work we've done the associated study of algorithms and ways for modelling the eruption of epidemic, with the main focus on influenza that should be prevented at an early stage if unfold. In the Introduction section we've given the importance to the study of relevant micro-blogging websites like Twitter, Face book, etc., that studies Social media platforms. In connected Work, we've done a survey from totally different resources and concepts applied to predict and notice the eruption of epidemics to study their benefits and limitations. Subsequently new plan is projected which may overcome the restrictions of models that are projected. This projected model contains machine learning technique to train the model, replacement plan for artiodactyl Epidemic Hint rule which can take care of epidemic activities happening on the Twitter and the Markov chain state model to categorise epidemic activities into 3 stages (Beginning of Epidemic, unfold of Epidemic, and Decay of Epidemic). Finally, we've projected a replacement framework to model epidemic prediction and KNN classification supports the scope of enhancements of previous work done.*

Keywords: *Twitter genus Apis, Markov chain State Model, BOWs, statistic classification, Knn, classification model .*

1. Introduction

Influenza is associate degree communicable disease caused by the contagion virus that spreads a speedy infection simply from one person to another person by coughing, instinctive reflex etc. It's a sort of variant virus that causes of 250000 to 500000 deaths worldwide per annum [1] [4] [5] [6]. It spreads supported multiple vectors like demographic, response to unwellness, etc. There are several disasters occurred within the past, the live examples show the results of a plague. In 1918 a plague named “Spanish Flu” caused the death of 50-100 million individuals.The impact of seasonal epidemics like the H1N1 is that the main anxiety of the health authorities. Though Researchers have created several medicines to regulate this, medicines simply ease the symptoms but doesnot cure the matter.As a result, it spreads terribly fast which makes extremely tough for elimination. Vaccination is alleged to be the foremost effective method to stop the infection of the contagion [3] on condition that, if the matter is detected early. The Social media plays a more robust role to inhibit and mitigate the hold of contagion like H1N1 as compared to the normal media. As a result of the social media which provides the important time data, however the normal media provides data once, something huge event happens that's not a defense. During this case even a tiny low delay is dangerous. As an example, if we rely solely on the normal media, the unwellnesss would unfold on quicker rate simply because of not detecting the epidemic at the proper time, that's very arduous to cure. Early discovery is just doable victimization using social media, i.e. micro-blogging websites like Twitter, Face book, MySpace. As social media plays a vital role in our day to day life, these platforms are foremost right channel for anyone to represent his/her opinion and feelings, by that he/she interacts with variant social users around the world, [9] As an example, health aware individuals share their diet plans on social media, not solely the diet plans like “How a lot of calories we have a tendency to burnt today” or “how to remain healthy”, they additionally share the way they fall sick as a result of some reason. Everyone seems to be sharing plenty of data associated with politics, culture, news and unwellnesss unfold. Within the different approach, it will be aforesaid that social media provides a virtual network that enable individuals to act with one another via the web. All data (status, tweets, etc.) hold on in social communities will facilitate to get the info associated with the epidemics. This data includes healthy (accurate) and false (inaccurate, inconsistent or insufficient) data. It's needed to make sure that the data or the knowledge given by the user ought to be correct and may be accustomed to stimulate positive health data and health outcome of individuals is improved instead

of false information that will threaten the general public safety [9]. Since data is power and powerful preventive measures will be taken before the unwellnesss unfold to the elements of the infected region, either within the style of instructed measures, victimizations, social communities or by operational health services. Twitter has become a well-liked medium for individuals to share their standing (tweets) in step with their mood, health problems, relationships, etc. These tweets also are updated by mobile phones by that we will trace the user's actual location and atmospheric condition. Twitter provides the free genus Apis from which a sampled read will simply be obtained. Twitter genus Apis enable North American nation to access only one sample of the Twitter information that is downloaded depends on the sampling technique we have a tendency to use. Twitter helps to make a map unfold model of unwellnesss. There are about 10 million active twitter users in India who often tweets via laptops, mobiles, iPods, etc., whereas collection of tweets may provides untried information and data supply through which we will extract the onset of contagion epidemics and its unfold [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 results of cdc|Center for unwellnesss management and Prevention|federal agency government agency|bureau|office|authority (Centers for Disease Control and prevention) information that is that the actual information, obtained by cases of contagion registered manually.

2. RELATED WORK

Jiangmiao et al. [1] provided a model to discover contagion transmissions. The model is predicated on "Sina Weibo", a Chinese micro-blogging web site that is sort of a hybrid of twitter and Facebook. They collected over 35.3 million tweets shared by all metropolitan cities in China. Knowledge was extracted on the idea of filtering techniques supported by federal agency ILI definition. They collected data associated with infection centre, the town set, target town, connected town and co-related town by mistreatment Dynamic Bayesian Network. They represented their results of detection and transmission at town level. In 2012, L. Bumsuk et al. [3] looked on tweets from twitter and compare the tweet corpus to Influenza- like unhealthiness (ILI) knowledge sets, weather factors, and also the contagion forecast. Comparison was created on the daily four-level contagion forecast from the peninsula meteoric Administration (KMA) and also the weekly ILI proportion peninsula Centre of malady interference (KDLC). In their results they represented comparison graphs of contagion signals on twitter to weather factors and contagion forecast. In 2011-2012, A. Harshvardhan et al. [4] [5] [6] Had planned a SNEFT design model that contained crawler, predictor and detector parts to predict respiratory illness activities mistreatment data gathered from micro-blogging websites like twitter and Facebook. During this framework, automotive vehicle Regressive Moving Average (ARMA) model was used to predict ILI incidences. This model worked with sure accuracy. Tools primarily utilized in this model were ARMA model, ARX model, OSN crawler. Tweets were collected from date 18th October 2009 to thirty first October 2010 and recorded 4.7 million tweets from 1.5 million distinctive users on their social relationships from twitter. Authors provided Hourly and weekly basis results by this model. Similarly, C.Aron [7] analyzed 500 million messages from twitter that took 8 months time and used filtering and regression. They obtained 95th correlations between their results and national health statistics. S. Takeshi et al. [8], the authors created associate earthquake coverage framework to discover earthquake activities. The framework explored the period of time nature of Twitter, specifically for event recognition. Linguistics analyses were utilized in tweets to characterize them in positive (tweets associated with the prevalence of the earthquake) and negative categories. The SVM (Support Vector Machine) that may be a machine learning algorithmic rule was used to train the information by giving positive and negative examples to the machine. Author planned a model that thought-about every Twitter shopper as a sensing element, and on those sensory observations, earthquake events were detected. To discover relevant activities space estimation ways, like Kalman filtering and particle filtering were used to quantify the areas of prevalence of the event. T.Xuning et al. [9] had planned a framework that was used to quantify users tormented by respiratory illness (swine flu) at intervals a social networking community associated introduced an UserRank algorithmic rule that incorporated the link structure, content similarity, responding order and time of repliers. They tested for flu forums that were of tiny size had 12 licensed members and every of them had 15.6 friends on the average. There have been 90 threads in total they tested and furnish their results with 100% preciseness L.Vasileios et al. [10], Authors planned a way for following the epidemic activity. Tweets on Twitter

within the UK over 5.5 million of users were ascertained so as to extract the calculations from twitter that measured the diffusion of ILI among the varied regions and tested the activities for 24 weeks and represented 95th accuracy compared to the official health reports of HPA (Health Protection Agency) and also the correlation coefficient table of twitter contagion score and HPA score and correlation graph comparison each. Recently, associate abundance of researchers are acting on the detection of the epidemics mistreatment using social communities like Twitter, Facebook, etc., to gather real time knowledge which will facilitate North American nation to stop and discover epidemics. Besides, several mathematical models are created, however still the area has some limitations, for instance, lack of real time data, machine learning tools, which will be used for predicting and sleuthing the epidemic.

3. Methodology

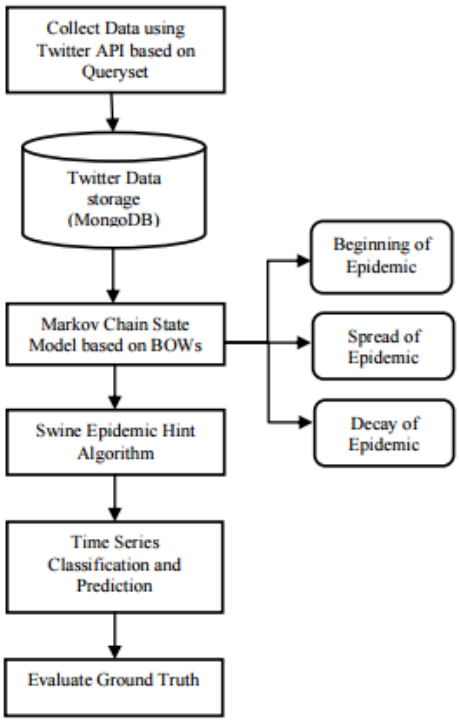
Till now, the work associated with the detection of the grippe epidemic, based mostly upon social networking communities that has been done on a awfully massive scale and totally different models even have been introduced. However, there's ample scope of improvement of those strategies attributable to the inherent nature of those strategies, accuracies and pertinence. Hence, the subsequent gap has been found as restricted work has exhausted the subsequent square measures: - A learning system ought to be trained so as to mechanically discover keywords that are additionally helpful to predict the bottom truth rate. - computer science may be utilized in combination with probabilistic models like Markov Chains rather than building grippe term corpus solely so model accuracy are going to be improved. - A model may be created which is able to work on a range of epidemics (if their symptoms square measure different) rather than one malady detection. - Work may be exhausted order to form model language freelance. - A model may be created which may notice and predict epidemic by considering several social media platforms KNN classification rule its classified supported class of symptoms.

4. Projected Work

After conducting systematic literature survey and learning material related to this drawback space, we tend to propose a brand new framework during which our model contains totally different techniques like machine learning algorithms and statistic classifications and prediction. This work may be accustomed to overcome the restrictions of the previous work done. During this model we are going to build a statistic classifications and prediction for identification of the epidemic stage based on probabilistic model of vocabulary (BOWs) utilized in totally different stages of the epidemic shared on-line by the act of tweeting.

5. Implementation

During this section we tend to picture the operation of projected work. The implementation of the projected framework is represented within the following steps:



A. Collect information from Twitter API supported question Set, once user tweets, the words used replicate the stage of epidemic unfold, these square measure supported the word/phrase/sentence connection score and also the stage of epidemic factors which will be accounted for the logic from detection of the epidemic begin. Hence, the question set that helps to extract relevant tweets and numerical factors helps to calculate the connection of text square measure thought-about as a feature set for doing the analysis work. Twitter's prevalence as a knowledge supply has prompted the advancement of applications and analysis in several areas. As Twitter used to remain

connected us with our followers and users, we have a tendency to follow and share tweets with one another through the globe, and it's additionally accustomed offer situational attention to an emergency circumstance. Several researchers have used Twitter to predict events like earthquakes and distinguish relevant shoppers to seem when to urge tragedy connected information. A sample of Twitter information will simply be obtained through the Apis that is freely accessible, to get the complete read is troublesome as a result of the Twitter Apis solely enable America to access 125th sample of the Twitter information, that is that the results of sampling strategy we'll use to associate with our needed data, whereas assembling Twitter information (runtime) a question set are going to be applied, that specifies a collection of keywords associated with swine flu Activities, in order that solely helpful data stores in sql.

B. Twitter information Storage (sql) when assembling the data (tweets) we've to store it somewhere that we will opt for any information. As an example, MS access, MySql, Oracle and sql. During this model, we select SQLdatabase as a result of its performance and fulfillment to the subsequent principles:

• Document-Oriented Storage- SQLstores its information in JSON-style objects, that makes it terribly simple to store raw documents from Twitter Apis.

• Index Support- SQLmakes it simple to form indexes optimized for our application, so it permits for indexes on any field

• Easy Queries- sql queries, whereas syntactically a lot of totally different from SQL, square measure semantically terribly similar. Additionally, SQLsupports MapReduce, that permits for simple lookups within the information.

C. Markov chain State Model supported BOWs Once, the gathering of Tweets info set is ample Bag Of Words for every Andrei Markov State(Beginning, unfold and Die) through that a pandemic undergoes. therefore Andrei Markov state model are often applied on information with regard to BOWs (Bag Of Words).These words, sentences, phrases, verb, adverb, noun verb, try mixtures show the state of affairs in terms of vocabulary tweeted by Twitter handler to its network which can extract the helpful content and can divide it into 3 states as given below-

• starting of Epidemic, this state can indicate the start stage of the epidemic.

• unfold of Epidemic, this state can narrate that the epidemic is unfolding or has already spread within the specific space.

• Decay of Epidemic, this state indicates that the epidemic in currently in restraint or died.

KNN classification:

During this algorithmic rule every relevant tweet text are going to be tokenized, and so slop words are going to be removed and last however not the smallest amount stemming also will be done. Once this step is complete, the numerical analysis of the tweet can begin supported the subsequent numerical formulas, which can check the connection of every tweet with regard to the state (Beginning of epidemic, unfold of epidemic and decay of the epidemic) tweet are going to be having.

BOWs square measure is a simplified illustration used for data retrieval [14]. The mix of words/sentences/keywords/phrases represents a Bag (multiset). The aim of Bag of Words is to retrieve those tweets containing keywords associated with even-toed ungulate epidemic activities with ease, speed and accuracy. We've 3 BOWs, each has its own significance, i.e. BBOW(Beginning Bag Of Words), has the keywords that indicate the start of the epidemic, SBOW(Spread Bag Of Words), keywords keep in it offer hint that epidemic is unfolding or has already spread and DBOW(Decay Bag Of Words), contains data regarding to the decay of the epidemic. Sort Score is that the Greedy Score during which “like operator” is employed for hard term frequency with regard to the bag of words for every Andrei Markov states of the epidemic . This operation essentially counts the words that square measure somewhat just like the words/phrases in BOW sets. However, just in case of the Equal score, actual sentence/word should match with BOW sets to search out term frequency. Tweets square measure divided into 2 classes one is “accurate” that has relevancy tweet and second is “not accurate” that isn't relevant. Preciseness is outlined because the relevant tweets retrieved by equal operator divided by the entire range of tweets retrieved by the search operator (like). Recall is outlined because the relevant tweets retrieved by Equal operator divided by the entire range of keywords comprised within the BOWs. The even-toed ungulate hint score is calculated at the tip that is that the final score of even-toed ungulate Epidemic Hint algorithmic rule declared mathematically because the product of preciseness and recall is split by the total of preciseness and recall i.e.

(Precision * Recall) / (Precision + Recall)

As mentioned earlier, we have a tendency to contemplate every of the characteristics in our training set as a unique dimension in some area, And take the worth an observation has for this

characteristic to be its coordinate therein dimension, and therefore obtaining a collection of points in area. We will then contemplate the similarity of two points to be, the distance and the area, the gap between them during this space underneath some acceptable metric.

The manner within which the algorithmic rule decides which of the points from the training set square measure similar enough to be thought-about once selecting the category to predict for a brand new observation is to choose the k nearest information points to the new observation, and to require the foremost common category among these. This can be why it's known as the k Nearest Neighbors algorithmic rule. A positive number k is fixed, alongside a brand new sample a pair of. We have a tendency to choose the k entries in our info that square measure nearest to the new sample three. We discover the foremost common classification of those entries four. This can be the classification we have a tendency to offer to the new sample

True class A (TA) - correctly classified into class A

False class A (FA) - incorrectly classified into class A

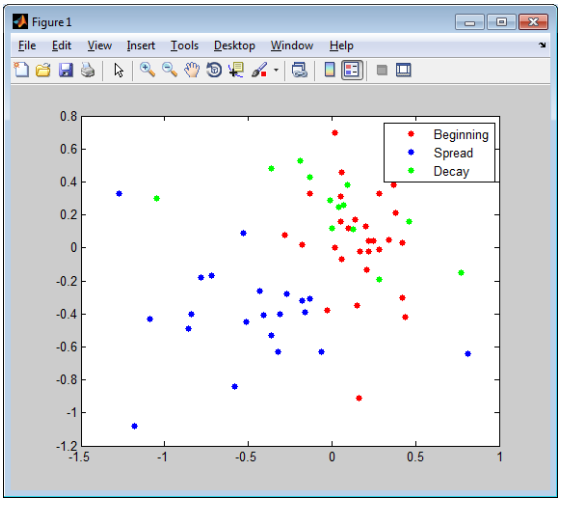
True class B (TB) - correctly classified into class B

False class B (FB) - incorrectly classified into class B

Precision=TA/(TA+FA)

recall = TA / (TA + FB)

True positive (TP), true negative (TN), false positive (FP), false negative (FN)



Statistic Classification and Prediction during this analysis, we have a tendency to work on the thought of Motif Recovery algorithmic rule for the statistic based classification of all the Andrei Markov states of the epidemic. Whereas mistreatment approach, machine learning algorithmic rule is additionally incorporated and machine learning algorithmic rule edges are often reaped, since our information (Tweet text) is time dependent. However, it should be done by removing the temporal ordering of individual inputs. Once the information is reworked, multi-linear regression algorithmic rule is also applied. the model can proceed for analysis and ground truth validation mistreatment Delphi technique [15].This would finally build put down rater agreements supported the scores given by the KNN algorithmic rule mechanically.

7. CONCLUSION

After doing the deep study of epidemic models and methodologies, we projected a brand new model that has not been used nevertheless and can overcome the drawbacks of previous work has been done before. This model can embrace machine learning formula within which system are trained so as to capture and react to grippe like activities so the preventive measures is taken to induce obviate the epidemic as early as potential. During this model we are going to use Twitter Apis to gather the tweets from the Twitter supported the question set. These extracted tweets are about to store in SQL (like the massive knowledge construct is there). Tweets collected from Twitter are additional classified in 3 Andrei Markov states (Beginning of the epidemic, unfold of the epidemic and decay of the epidemic). On these Andrei Markov States, a artiodactyl mammal Epidemic Hint formula are applied for convincing the score of the tweet. At the tip Ground Truth are evaluated victimization the Delphi technique which might finally build Associate in Nursing repose rater agreement supported the scores given by the artiodactyl mammal Epidemic Hint formula mechanically. As per our estimation this model can offer higher accuracy and can facilitate within the prediction, detection and management the unfold of disasters that are available the long run.

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.