International Journal of Advanced Science and Research
ISSN: 2455-4227
Impact Factor: RJIF 5.12
www.allsciencejournal.com
Volume 3; Issue 2; March 2018; Page No. 19-24

Sentiment analysis using machine learning techniques to predict outbreaks and epidemics

[1] Rameshwer Singh, [2] Dr. Rajeshwar Singh, [3] Ajay Bhatia
[1, 3] Research Scholar, I.K. Gujral Punjab Technical University, Jalandhar, Punjab, India
[2] Group Director, Doaba Khalsa Trust Group of Institutes, SBS Nagar, Punjab, India

Abstract
Sentiment Analysis (SA) is the study of the datasets available over web that contains fruitful information. Machine learning techniques have a great role in computer science. One of the applications of machine learning techniques is to study patterns and development of such computational systems that can learn themselves and make predictions. In recent past there are examples of studies in which data sets are taken from microblogging websites like Twitter etc. and significant healthcare information generated by sentiment analysis of data sets. In this research article, (i) significance of sentiment analysis in predictions is described; (ii) relevance of sentiment analysis using Machine Learning (ML) techniques is studied; (iii) classification of data based on ML techniques is also described; (iv) survey on use of microblogging sites to predict the outbreaks and epidemics is done through some major research articles from year 2010 to 2017.

Keywords: sentiment analysis, machine learning, outbreak prediction, twitter data mining

1. Introduction
One of the applications of machine learning techniques is to study pattern recognition and development of such computational systems that can learn themselves. The machine learning algorithms, train the system on the basis of available training data sets and after the training the system can predict the future data values.

Epidemic means when a disease rapidly spreads in huge population within a short period of time. Generally, when large number of people got infected by a disease spreading with in a country or the part of the country is called epidemic [1]. On the other side, if the disease is spreading in countries or continents then it is termed as pandemic. There can be two types of outbreaks- Common source outbreak which will affect the population with common agent and the other type is Propagated outbreak i.e., the disease which travels from person to person [2, 38].

There is a worldwide long history of epidemics since 429-426 BC to till date. Starting with the most recent ones- year 2017 Japanese encephalitis in India, 2015 Swine flu outbreak in India, since 2014 Odisha jaundice outbreak in India, since 2013 Ebola outbreak worldwide, since 2011 Dengue outbreak in Pakistan, Since 2011 hand, foot and mouth diseases in Vietnam, since 2011 measles in Congo, cholera at Hispaniola since 2010, influenza in 2009, and many other diseases across the world like Meningitis, mumps, hepatitis B, bubonic plague, malaria, yellow fever, HIV/AIDS [3, 4] etc. So, epidemics are the severe problems worldwide.

Especially in countries like India, where there is lots of diversity in terms of nature, location, people etc. There is always a big threat that an infectious disease can convert into epidemic. Considering the past epidemics like smallpox in 1974, plague in 1994, in 2009 flu and hepatitis, in 2014 Odisha jaundice and Swine flu in 2015 that are faced by India there is a high need for deep study on these kinds of epidemics [3, 5].

As discussed above, a disease turns into epidemic because of its infectious nature which made it travel from person to person or by other mode into the huge amount of population. The study on epidemic patterns will help in many ways. It will provide efficient ways to restrict epidemic from being spreading at more locations. It can also help to improve the survival rate, recovery rate and decrease the death rate by processing the useful data that is being shared by many people on social networking sites and other publically available data. The processing of such data can generate important information that is beneficial for health care to minimize the economic and human loss of a country [22, 38].

To better understand the epidemic patterns, the machine learning based computational model is required that can predict the valuable information on time. The technology can help to contribute in terms to find the epidemic patterns and a support to control it a genuine manner. It has been presented in so many research articles that many outbreaks could have been controlled on time if we had listened to data on social media and other publically available data [26, 27]. The technology like machine learning, intelligent system, natural language processing can contribute in this regard by generating the valuable information on time [38].

2. Machine learning techniques in sentiment analysis
2.1 Machine learning techniques
Machine learning techniques in recent era are very useful to make automating classification, clustering and predictions. Machine learning techniques in most of the cases have data sets for training and data sets for testing. With the help of training data sets system learns how to classify the test data sets, by analyzing its classification one can make future

decisions [37].

2.1.1 Supervised Learning
Supervised learning gets training from existing labeled data and used for classification of data. In supervised learning algorithms map the function's input to respective outputs. If the outputs belong to a particular class, it is known as classification otherwise it will be called regression problem.

2.1.2 Unsupervised learning
Unsupervised learning doesn't have labeled data sets for training for clustering. In unsupervised learning, training task is done through the inputs directly. Structure and relation between the inputs are automatically identified by the learning algorithm. It is appropriate for the clustering of inputs to put them into appropriate cluster.

2.1.3 Semi-supervised learning
Semi-Supervised learning uses both labeled and un-labeled data for learning [36]. Semi supervised learning the branch of supervised learning techniques. As mentioned these techniques shall use labeled and unlabeled data. In these techniques minimum labeled data is used as training data sets. Hence is overcome the problem of preparing the large training data sets with labels that are required in case of supervised learning techniques.

2.2 Popular machine learning algorithms
2.2.1 Nave Bayes
Nave Bayes uses all the features of feature matrix. All these features are studies separately as individual feature and estimates the probabilities in prior [36, 37]. In article [33, 34, 35] use of Nave Bayes is presented. In general nave bayes is a probabilistic classifier based on bayes theorem.

2.2.2 Support vector machine
Support Vector Machine finds a linear separator that can classify the given data in the best possible manner [36, 37]. Article [19, 33, 35] presents the successful implementation of the SVM in predicting the outbreaks. SVM is one of the good classifier that classifies by separating the hyperplane.

2.2.3 Maximum entropy
In Maximum Entropy, no assumptions are taken between regarding the features relation with each other. Maximization of entropy is the major objective of this technique [37]. In general, it is states the probabilistic distribution that represents the state of knowledge with highest entropy.

2.2.4 Decision Tree
Decision Tree makes the samples of training data based on the features. It classifies the data based on hierarchical decay [36]. Article [24, 33, 34] present the use of decision trees.

3. Review of literature
Sentiment analysis is the feasible area of research identified by researchers to find the relation between sentiment of people and its relation with actual facts. SA using ML based techniques is helpful to make predictions.
The survey of SA using ML techniques to predict epidemics and health care information of article covers a decade from 2004 to 2016. This survey contributes how SA is useful for prediction of generic and health care information. Review of literature is classified into three parts: Use of ML techniques in SA, Classification and predictions using SA, Prediction of epidemic and health care information using public data.

3.1 Sentiment analysis using machine learning techniques
In survey paper [6], Authors presented Sentiment Analysis as the study of sentiments, emotions of the people on some event, entity or anything. The process used in paper has phases like Product Reviews - Sentiment Identification - Feature Selection - Sentiment Classification - Sentimental Polarity. The paper presented two broad categories: 1. On the bases of machine learning techniques. 2. Lexicon based approach. There were different classifiers which could be used like decision tree classifier, linear classifier, rule based classifier and probabilistic classifier in case of machine learning based approach [38].
In the article [7], author had examined the practice of linguistic characters to perceive the sentiments of twitter message. Three different corpora of twitter message were used: hashtagged based, Emoticon based and iSieve Corporation data sets. Pre-processing on data was done i.e., tokenization, normalization, part-of-speech (POS) tagging. The outcome of experiments was that the part-of-speech features were not useful for sentimental analysis in microblogging domain. Microblogging features like positive/negative/neutral and others were most useful [38].
In paper [8], authors presented lexical based and machine learning based approaches. Comparisons of both the techniques were also based on movie reviews. The five different variants of lexical approach were discussed. On the other side in case of Machine learning approaches apply the Part of Speech to each blog post and create a set, after that traverse all the posts in experimental set to find number of positive, negative and uncategorized words. The experimental results stated that machine learning approach was having very high accuracy as compared to the lexical based approach. Similarly, in paper [9], movie review based study presented using Naive Bayes machine learning [38].
In the article [10], presented three approaches for sentiment analysis: -1. Machine learning based approach. Machine Learning Techniques like maximum entropy, Naive Bayes, and support vector machines already have great success in the field of sentimental analysis. 2. Lexicon based approach in which prior train data sets were not required and classification was done via comparison of the features. 3. Hybrid approach, which combined the best features of both. In the comparison study of the above mentioned techniques although machine learning had highest accuracy but it required large training sets. On the other side Lexicon based approach works for short sentence and gives better performance. In case of hybrid approach best features of both the models can be used [38].

3.2 Classification & prediction using sentiment analysis
In paper [12], authors shared the idea about how we could predict the future based on the content available on the social networking sites. Chatter on twitter regarding forecasting the box office revenues for movies were taken in account. A

simple model was proposed which demonstrated the use of tweets to make market based predictions. Linear regression model was used to predict the box office revenues. There was high correlation between the amount of attention given by people and ranking in the future [38].

In research article [13], classification of tweets was done through two models: one was binary classification, second was 3-way classification. The experimentation was done through three prototypes unigram, feature based and tree kernel model. Among the mentioned three models tree kernel based model performed better than other two models. In the paper classification of micro blogging data, datasets, pre-processing techniques for data, polarity schemes, design of tree kernel, feature based approach and experimentation on data were discussed [38].

In the paper [14], an algorithm was developed to analyze the emotional polarity automatically. Initially combined approach was used using SVM and K-means with objective develop unsupervised mining approach. Before the cleansing of data total 510,218 posts were collected and after cleansing of the data remaining number of posts were 220,053 from 31 forums. Using K-means 52 clusters were generated. The deficiency in the k-means was that on pre-determined value of k was required. The drawback can be overcome taking the range of k between 5 to 20. During experimentation SVM, binary based classification was used to identify the hotspot among 31 forums. Five metrics sensitivity, correctness, specificity, positive predictive value and negative predictive value were used to make comparisons between K-means and SVM. The experimental result showed that results obtained using the SVM were more reliable [38].

In [15], authors presented a system that was used to perform real time analysis of public sentiments for the president candidate in the 2012 U.S elections. The sentimental model used in this paper was based on assumptions that opinion may be highly subjective or contextualized. The train data in this paper consisted of 17000 tweets out of which 56% were negative, 16% were positive, 18% were neutral and remaining were unclear [38].

Another article [16], presented the public opinion about the transportation system using data mining of the data from social networking websites. In this paper twitter tweets were analyzed regarding the about light rail transit services in Los Angeles [38].

In paper [28], presented motion tracking and how to learn about different patterns based on the observations. A novel, probabilistic method is shown. The proposed system was able to track the people in indoor and outdoor environment [38].

In review article [29], stated different issues involved in the pattern analysis like feature extraction, selection, ordering, regression, grouping, and validation etc. The objective of the study is to summarize the most widely techniques used for the pattern analysis and also highlighting the research in this era [38].

In article [30], review on machine learning techniques for handling the irrelevant information is presented. A general framework is presented to compare the different methods [38].

3.3 Epidemic & outbreak predictions

In paper [19], authors proposed a system that catch the influenza related tweets after that the actual influenza patients were mined using the SVM based classifier. The data extracted from the twitter had 42% negative tweets that included the word "influenza". The results showed very high correlation of 0.89. In the train data sets there were about 5,000 tweets of November 2008 and test data sets are other. In case of excessive news periods the methods not worked that well because of bias news. However, tweets have great advantage especially in case of early detection [38].

In article [20], proposed a dengue surveillance approach which weekly overviews and compare it with a week before data. Twitter data grabbed with four dimensions that were: Volume, Location, Time, and Perception. Tweets were divided into five categories as Personal experience, Ironic/sarcastic tweets, Opinion, Resource and Marketing. Selective sampling was done to make the train set after that correlation analysis is done. Spatio-temporal analysis was useful to detect the epidemic in early stages. Data sets used in this article are official dengue reports and twitter messages [38].

In paper [22], application of twitter for public health discussed. A model was applied nearly on one and half million health associated tweets, with the objective to track the illness over the time, measuring the behavior risk factors, symptom of disease, medication usage etc. The results showed that twitter have great applicability in the research area public health [38].

In [23], authors proposed several methods to identify influenza related messages and its correlation with CDC statistics with the help of different regression models [38].

In paper [24], authors presented the influenza detection algorithm that automatically differentiates relevant techniques. The estimated influenza frequency of the system was correlated with the data of health and mental hygiene of New York City and Centers for Disease Control. The proposed system provides 85% accuracy. In this paper supervised classification techniques were used. Algorithm proposed in the paper had shown important enhancements and less sensitive to the twitter users by concentrating on the reports of influenza infection [38].

In research paper [25], a system SNEFT was proposed that will keep the track of messages on twitter related to flu and forecast the emergency and influenza in the population. The authors found high correlation between the data of ILI (influenza-like illness) cases informed by CDC and the flu related tweets posted during the 2009 and 2010. The paper also proposed an auto regression model to predict the level of ILI movement in population. The experimentation was done on CDC data with twitter data and without twitter data. Results showed that twitter data helped in achieving high accuracy [38].

In paper [26], objective was to monitor the use of term H1N1 vs Swine Flu, to perform analysis on tweets and to confirm the use of twitter as a tool for sentiment analysis. The study done in the article states that the H1N1 related tweets were primarily shared to spread information from the reliable source. Further these tweets can also be used for real time sentiment analysis to spread awareness in health [38].

In article [27], authors presented the use of twitter in predicting the outbreak of swine flu in 2009. The findings of the paper were that the twitter contains lots of spam, noisy data so a spam filter can be used to avoid this limitation. The use of

twitter would be beneficial in an epidemic intelligence system [38].

In research article [32], Swine Flu Hint Algorithm is proposed The base of the study is time series based classification and predictions. N1H1 surveillance system is proposed in article [33], RSS feeds of different news agencies and tweets were used as data sets. Classification of data is done suing various techniques like Naïve Bayes, SVM, Random Forest and Decision Tree. SVM performed better than other techniques. In article [34] a surveillance system is proposed for Dengue in Malaysia. In this article data related to Weather, Food, Social Media, Dengue Symptoms and Build Environment was used. Large data set was analyzed using machine learning techniques like Functional Trees, Bayesian Logistic Regression to find the symptoms of dengue. The proposed system helped in detecting the dengue outbreak at early stages. Article [35], suggested the importance of geo-tagging in the prediction of misquote related diseases. Proposed system is based upon on Naïve Bayes and Support Vector Machine. Twitter and RSS feeds are used for the source of data. Article proved that machine learning techniques are better than traditional approaches.

Table 1: Research article reviewed on use of social media to predict outbreaks and epidemics

Ref .No	Type	Publication Details	Technique(s)	Data Source
19	Flu Detection Through Twitter	." Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011.	Support Vector Machine	Twitter
20	Dengue surveillance approach	." Proceedings of the 3rd International Web Science Conference. ACM, 2011	Spatio-temporal analysis	Twitter and Official Dengue Reports
22	Twitter to analyze public health	ICWSM, 2011	Ailment Topic Aspect Model	Twitter
23	Detecting Influenza Epidemic through Twitter	Proceedings of the first workshop on social media analytics. ACM, 2010	Simple Regression Model	Twitter and CDC
24	National and Local Influenza Surveillance using Twitter	PloS, 2013	Supervised Classification Model	Health and mental hygiene of New York City, Centers for Disease Control and Twitter.
25	Prediction of Flu Trends using Twitter	IEEE, 2011	ARX Model	CDC, Twitter
26	Monitor the use of term H1N1 vs Swine Flu	PloS, 2010	Automatic Query System	Twitter & RSS Feeds
27	Prediction Outbreak of Swine Flu through Twitter	Springer Berlin Heidelberg, 2012	Normalized cross-correlation ratio between various signals from Twitter and the official HPA surveillance data	Twitter and UK surveillance data
32	Influenza prediction model.	IEEE, 2015	Time Series Based Classification and Prediction	Twitter
33	Tracking of N1H1 Pandemic in India	Procedia Computer Science, 2015	Naïve Bayes, SVM, Random Forest and Decision Tree Classification Techniques	Twitter, New Feeds (e.g. TOI, IBN, Zee News etc.)
34	Surveillance system for Dengue in Malaysia.	IEEE, 2016	Functional Trees, Bayesian Logistic Regression	Data related to Weather, Food, Dengue, Build Enviroment from social media i.e. Twitter, Facebook etc.
35	Surveillance and predictive mapping of misquote based diseases using social media.	Journal of Computational Science, 2017	Naïve Bayes and Support Vector Machine	Twitter and RSS Feeds

4. Conclusion

In the recent years, there are lot many studies done on the use of Machine Learning Techniques to predict the Epidemics and Outbreaks. Many examples of successful applicability of Twitter and other social media data in predicting the different outbreaks or the relation between the microblogging sites and the real data presented in the article. However, it seems that there are some deficiencies, which needs the attention to produce better result through machine learning based sentiment analysis. Different machine learning techniques mentioned above have their own advantages and drawbacks. There is a need to propose such a model that describe the best ways of data collection, data filtering and machine learning techniques that can help in predicting the epidemics at early stages.

In thus paper almost research work done in one decade is presented with main focus on the predictions of health care information, outbreaks and epidemics using machine learning with data sets available on social media, RSS News Fees and data from health care agencies. The results obtained in each result were quite satisfactory and motivate their application in future and other domains. All the techniques and studies that are discussed in the article can be used foundation for further research work.

5. References

1. Martin Paul MV. Estelle Martin-Granel. 2,500-year evolution of the term epidemic. Emerging infectious

diseases, 2006; 12(6):976.
2. Taylor Ian, John Knowelden. Principles of Epidemiology. Principles of Epidemiology. Edn., 1964, 2.
3. Wikipedia. List of Epidemics. URL http://en.wikipedia.org/wiki/List_of_epidemics (accessed on 1st June, 2015), 2015.
4. CNN International Edition. URL: http://edition.cnn.com/interactive/2014/10/health/epidemics-through-history/ (accessed on 30th May, 2015).
5. India struggles with deadly swine flu outbreak. BBC News. Retrieved 21 February 2015. URL: http://www.bbc.com/news/world-asia-india-31547455 (accessed on 30th May, 2015), 2015.
6. Medhat Walaa, Ahmed Hassan, Hoda Korashy. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal. 2014; 5(4):1093-1113.
7. Kouloumpis Efthymios, Theresa Wilson, Johanna Moore. Twitter sentiment analysis: The good the bad and the omg!. ICWSM. 2011; 11:538-541.
8. Annett Michelle, Grzegorz Kondrak. A comparison of sentiment analysis techniques: Polarizing movie blogs. Advances in artificial intelligence. Springer Berlin Heidelberg. 2008; pp. 25-35.
9. Kumari Pooja, *et al*. Sentiment Analysis of Tweets. IJSTE, 2015.
10. Kharche Ms Swapna R, Lokesh Bijole. Review on Sentiment Analysis of Twitter Data. International Journal of Computer Science and Applications, 2015; 8(2).
11. Pang Bo, Lillian Lee. Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2008; 2(1-2):1-135.
12. Asur Sitaram, Bernardo A. Huberman. Predicting the future with social media. Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. IEEE, 2010, 1.
13. Agarwal Apoorv, *et al*. Sentiment analysis of twitter data. Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics, 2011.
14. Li Nan, Desheng Dash Wu. Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Support Systems. 2010; 48(2):354-368.
15. Wang Hao, *et al*. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012.
16. Luong Thuy TB, Douglas Houston. Public opinions of light rail service in Los Angeles, an analysis using Twitter data. iConference 2015 Proceedings(2015).
17. Raut Dhanashree, Seema Ladhe. An Empirical Approach for Semi-Supervised Sentiment Analysis and Opinion Mining.
18. Tan Songbo, Jin Zhang. An empirical study of sentiment analysis for chinese documents." Expert Systems with Applications. 2008; 34(4):2622-2629.
19. Aramaki Eiji, Sachiko Maskawa, Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using Twitter. Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011.
20. Gomide Janaína, *et al*. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. Proceedings of the 3rd International Web Science Conference. ACM, 2011.
21. Buscaldi Davide, Irazú Hernandez-Farias. Sentiment Analysis on Microblogs for Natural Disasters Management: a Study on the 2014 Genoa Floodings. Proceedings of the 24th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2015.
22. Paul Michael J, Mark Dredze. You are what you Tweet: Analyzing Twitter for public health. ICWSM, 2011.
23. Culotta Aron. Towards detecting influenza epidemics by analyzing Twitter messages. Proceedings of the first workshop on social media analytics. ACM, 2010.
24. Broniatowski David A, Michael J Paul, Mark Dredze. National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. PloS one. 2013; 8(12):e83672.
25. Achrekar Harshavardhan, *et al*. Predicting flu trends using twitter data. Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on. IEEE, 2011.
26. Chew Cynthia, Gunther Eysenbach. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PloS one. 2010; 5(11):e14118.
27. Szomszor Martin, Patty Kostkova, Ed De Quincey. # swineflu: Twitter predicts swine flu outbreak in 2009. Electronic Healthcare. Springer Berlin Heidelberg, 2012, 18-26.
28. Stauffer Chris, Eric LW Grimson. Learning patterns of activity using real-time tracking. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2000; 22(8):747-757.
29. Gutierrez-Osuna Ricardo. Pattern analysis for machine olfaction: a review. Sensors Journal, IEEE. 2002; 2(3):189-202.
30. Blum Avrim L, Pat Langley. Selection of relevant features and examples in machine learning. Artificial intelligence. 1997; 97(1):245-271.
31. Shawe-Taylor John, Nello Cristianini. Kernel methods for pattern analysis. Cambridge university press, 2004.
32. Grover Sangeeta, Gagangeet Singh Aujla. Twitter data based prediction model for influenza epidemic. Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on. IEEE, 2015.
33. Jain Vinay Kumar, Shishir Kumar. An Effective Approach to Track Levels of Influenza-A (H1n1) Pandemic in India Using Twitter. Procedia Computer Science. 2015; 70:801-807.
34. Othman Mohd Khalit, Mohd Shahrul Nizam Mohd Danuri. Proposed conceptual framework of Dengue Active Surveillance System (DASS) in Malaysia. Information and Communication Technology (ICICTM), International Conference on. IEEE, 2016.
35. Jain Vinay Kumar, Shishir Kumar. Effective surveillance and predictive mapping of mosquito-borne diseases using social media. Journal of Computational Science, 2017.
36. Aydoğan Ebru, Ali Akcayol M. A comprehensive survey

for sentiment analysis tasks using machine learning techniques. INnovations in Intelligent SysTems and Applications (INISTA), 2016 International Symposium on. IEEE, 2016.
37. Neethu MS, Rajasree R. Sentiment analysis in twitter using machine learning techniques. Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.
38. Rameshwer Singh. Ph.D Synopsis Submitted to IKG Punjab Technical University, Jalandhar (Unpublished), 2016.