International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Sentimental Analysis of Twitter Health Data using Machine Learning Techniques

E.M.Roopa Devi, R.Rajadevi, S.Vinoth Kumar

Abstract: ***With the huge development of Internet, more users have occupied with wellbeing networks, for example, medicinal discussions to assemble wellbeing related data, to share encounters about medications, treatments, analysis or to associate with different clients with comparable condition in social media. A lot of lookup has focused on examining Twitter health tweets for subject matter modeling using quite a number clustering approaches, but few have mentioned it for sentiment analysis. The truth that such statistics carries potential information for revealing the opinion of humans about fitness services and behaviors make it an interesting study. In these paper, universal sentiments about Twitter health data was investigated. Twitter, measuring and monitoring the occurrence of social health problems. The approach is based on two stages: In first stage separating perhaps applicable tweets utilizing a lot of uniquely made standard articulations, and afterward arranges these underlying messages utilizing machine learning techniques. Using the Twitter search API and Twitter metadata geocoded content, social media tweets were selected to start filtering. Once Tweets are correctly identified, the classifier was applied to data in order to filter out the tweets. Classification results were improved by detecting the values of ROC and f-measure. This report indicates that such a method provides a viable solution for quantifying and tracking the progression of health status within society.***

Keywords :Twitter health information, Sentimental Analysis, Feature selection, Machine learning Techniques.

I. INTRODUCTION

Twitter is an online life stage the spot more than 500 million individuals worldwide presents their thoughts and work out different points, which incorporate their wellness stipulations and open wellness occasions.In view of the number of facts exchanged by both people and reputable sources, Twitter has proven to be a required supply of health statistics on the Internet. Twitter records have been described as useful to a number of public health applications, including:(1) disease tracking, (2) public response, (3) emergency / emergency situations, (4) forecasting, (5) lifestyle, (6) geolocation, and (7) regular applications. Certifiable wellbeing administrations and practices can be advanced using Twitter wellbeing news.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
E.M.Roopa Devi*,Information Technology, Kongu Engineering College, Perundurai, India.
R.Rajadevi ,Information Technology, Kongu Engineering College, Perundurai, India.
S.Vinoth Kumar, Information Technology, Kongu Engineering College, Perundurai, India.

In contrast to a large portion of the ecologic appraisal strategies which catch and enable individual to report current area, movement and social encompassing at any specific period (Schwartz and Stone, 1998), tweets are not explicit irregular improvement reliant as they speak to increasingly naturalistic substance with the accessibility in huge volume as an additional preferred position (Yoon et al., 2013). Performing substance mining investigation via web-based networking media information so as to concentrate individuals' feeling about wellbeing administrations and practices is significant in light of the fact that picking up full comprehension of such conclusions from content has been troublesome because of their multifaceted nature. Web substance mining centers around finding significant information (for example themes, estimations) from information, for example, web journals, web based mailing records, and internet based life by applying different methods, for example, AI, information mining, data recovery, regular language handling and measurements The wealth of data accessible in web based life and wellbeing news discussion alongside free and rich articulation of conclusions has pulled in general wellbeing network to dissect the response of individuals to wellbeing administrations and practices (Nasraoui, 2008).The procedure of consequently estimating the theory, feelings, assessments, and supposition communicated in a content is alluded to as sentimental analysis. The theory, feelings, assessments, and sentiment can either be negative, positive or impartial. The article's description is as follows. Update on related work on gathering social media health information in Section 2. Discuss the methodologies for the identification of twitter feelings in section 3.The results of the experiment on various twitter health data in section 4.Discuss the premise and potential research in section 5.

II. RELATED WORK

The motive of applying opinion examination strategies is to prepare wellbeing related conclusions of millions of users and conclude them towards valuable data. Hence, the result of opinion examination needs to be basic and conclusive that can be utilized for the reason of decision making. Paul et al. proposed a new model of associative topics to identify ailments tweets. This model, called the Ailment Topic Aspect Model (ATAM), uses a combination of keywords and associated topics to identify relevant tweets. Denecke et al. presented a prototype implementation of a ailment surveillance system known as M-Eco that strategies social tweet records for significant disorder outbreak data.

Bosley et al. have analyzed and classified 60,000 tweets of cardiac arrest and revivification received over a 38-day period using a seven-phrase collection.. Behera and Eluri suggested a sentiment analysis approach to track disease transmission by location and time. Salas-Zarate *et al. have introduced an* aspect-level approach for sentiment analysis on tweets about diabetes. Missier et al. reviewed two different approaches to identifying dengue-related Twitter data and other outbreaks of Aedes-borne disease. Myslín et al. have used twitter data to test the public's perception of smoking and tobacco related products. Adrover et al. have analysed to identify Twitter users with HIV and decide whether Twitter could track drug treatments and their related feelings. Ji et al. used Twitter to monitor public disease spread. Their techniques included dividing tweets into private and (non-personal) news groups to focus on public affairs. Schulz et al. proposed a dual-label training method for analyzing the classification of event-related twitter information. Al-Amrani et al. have introduced a combination of random forest and vector support system to derive opinion from the data set of product review. Ribeiro et.al suggested a coherent approach to sentiment analysis, which consists of four components: collection of data, noise elimination, generation of lexicons, and identification of sentiments. THIS works illustrate that use of Twitter information to analyze a broad range of health issues. Lexicon-based approach involves conducting report and sentence level sentiment analysis by looking for word polarity from the predefined word list. The tweets are marked manually as positively or negatively or unsure for learning classifiers to identify the state of health. In order to start filtering, social media tweets are collected using geocoded information from the Twitter search API and Twitter metadata. In this analysis they merged a system of lexical analysis with a process of identification to identify appropriate references to the occurrence of disease and other conditions of health.

III. PROPOSED WORK

In this research, propose a tool for gathering a series of tweets that indicates the presence of certain human health conditions as a way of inferring the prevalence of such conditions in society. Figure 1. show the process of strategies that they use to improve the performance of tweet health dataset.

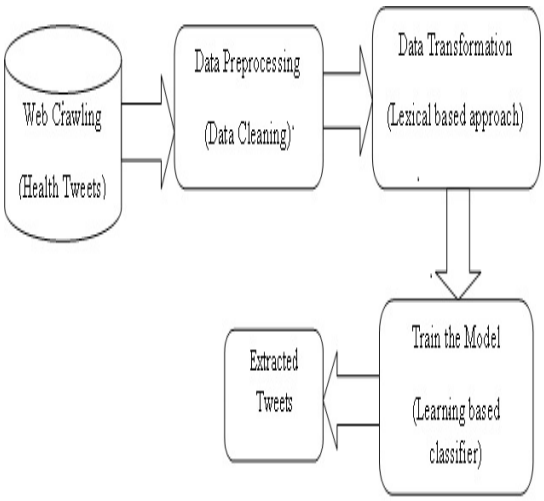

Fig.1. Schematic Diagram for Sentimental Analysis

The proposed work of sentimental analysis on health data include the following procedure

A. Data Collection

The data collection is carried out via the Twitter API. The Twitter API allows users to communicate with their information, i.e. tweets. Through creating a twitter API, users can download such tweets. The user requests data from the API and returns data according to the application query.

B. Data Preprocessing

Twitter information consists of noisy records such as RT for Retweets,' #' hashtags for the topic-based sorting of tweets, @usernames, external net links, and emoticons. Preprocessing feature eliminates all noisy data so that a fact is smooth and operations on easy data are easy to perform. 1) Delete Duplicate tweets; 2) Eliminate Retweets; 3) Remove URLs; 4) Delete Unnecessary Space; and 5) Remove twitter hashtag.

C. Lexical Based Approach

Lexicon Based Sentiment Analysis is concerned with the inclusion in the text of certain terms. Lexicon contains various features including the word marking component, the principles of their emotions, word subjectivity, etc. Tweet Sentiment Analysis is annotated using these lexicon features. By averaging the sentiment values of words, it can achieve polarity of the whole tweet. Negation Handling is significant issue while opinion investigation. Since many sentence contain the invalidation word that moves the extremity of the sentence. Numerous classifiers expel the repudiation words by thinking about it as stop words

D. Data Transformation

To construct such expression, a collection of tweets containing the name of each condition was first collected, tweet replies and tweets containing references were removed and the log probability of words within those datasets were determined, creating an ordered list of words identified with each condition. Tokenization was performed in the tweet transformation stage to represent tweet content as a function vector. All cases are eventually turned into smaller cases. Tokenization and word-case conversion were carried out using RapidMiner operators "tokenize" and "transform cases"

E. Sentence Type Detection

The use of statistical computational techniques, emotional analysis and opinion mining provides valuable insight into the author's perspective or emotion of a document or subject by exposing whether the voice is negative, optimistic or neutral. Subjective tweets are the tweets that contain the sentiment of the recipient, the perception of something in the world. Therefore, it is necessary to find the subjective tweets to identify tweets as Subjective and Objective tweets. The text is objective(To transmit factual information on the environment).or subjective (Reflecting the feelings and beliefs of the author)

F. Polarity Detection

Tweet polarity by looking for word occurrences in the lexicon dictionary and simply replacing the word location with the polarity meaning shown in the lexicon dictionary.

The polarity of the entire tweet depends on the sum of the word polarity of the tweet. Check for the occurrence of each tweet word in a lexicon dictionary when it is found to substitute the word with the lexicon's polarity quality. Notice the term does not exist in the lexicon and then substitute it with the polarity value zero showing the neutral polarity. Aggregate all word polarity values in tweets that indicate the tweet's polarity with the aggregate value. These tweets are the training data used to train classifiers.

G. Machine Learning

The polarity of the lexicon dictionary can be viewed as training data for each sentence. To train the classifier, these training data are given to the Machine Learning Classifier. To measure polarity of other data by training using this training data that can be transferred to the classifier as test data. Make use of regular expressions in huge sets of tweets relating to the medical problem and medication specified. Negative phrases that do not reflect an individual's presence or absence of a disorder in one person, such as " Hoping cough will hit me again this winter " may also appear among the tweets received. To resolve this issue, using a Machine learning methods use regular expressions to process the data set to filter the tweets.. To apply machine learning, a set of features will reflect each tweet. After eliminating stopwords and word stemming, using a simple bag of words template. Trialing the appropriate classifier for our model using different machine learning techniques such as (Support Vector Machines (SVM), Naive Bayes (NB). The distinct sets of tweets were classified before the experiments were carried out. For every tweet is analyses as positive, negative or uncertain.

- Positive: The tweet data refers to presence of one of the diseases / states studied in the person who wrote the tweet.
- Negative: The blog substance demonstrates that no one of the infections / states is considered by the individual who composed it.
- Uncertain: The tweet message does not comply with any of the above specifications

H. Feature Selection

There are typically some redundant or obsolete features in the bag of words and characters n gram applications. A sub-set of relevant features is selected and used in this step to construct the model of classification. The selection process of the feature give the following benefits: It eliminates distracting and redundant features. This distraction makes it difficult to find relevant patterns. In order to find appropriate classification patterns, larger training data sets are needed. However, the trained data collection is very small in a number of data mining applications. In this case, it may be useful to pick the role of using the models are based on a minimal number of features. It improves the classifiers ' effectiveness. The selection process involves two components: the features evaluator, the algorithm that decides the value of the subsets of features to be allocated to the class label; and the search tool, the method used to search for suitable feature subsets. As this issue increases with that of the sets of features, there are many ways to check the associated subsets efficiently.

Correlation based Feature Selection (CFS):It rank the attributes based on a correlation-based heuristic evaluation method. The method analyzes vector subsets of attributes related to a category tag however independent of one another. This approach implies that insignificant features have a low correlation class value and hence the features will be discarded. Excess number of features, at the other side, should be examined since they are usually highly correlated to one or more of the other features. One can describe the criteria used it to test the subset of x characteristics as stated:

$$Ns = \frac{\overleftarrow{xtcf}}{\sqrt{x}+x(x-1)tff} \qquad (1)$$

Where a subset of features Ns is evaluated in equation (1), t_{cf} is average value of correlation is among the class labels and features and t_{ff} the average value of correlation among the two feature is evaluated.

IV RESULT AND DISCUSSION

Setting up a benchmark by observing the feelings conveyed in medical blog posts. Emphasis on fine-grained aspects of the health status and treatment of users. Our intention is to enable the annotation scheme to capture multiple user health status perspectives. Two important medical elements with potential sentiment values categories:

Table- I Categorization of sentiment values

Section	Aspect1	Aspect2	Aspect3
Complication	Existence	Restore	Depreciate
Therapeutic	Fruitful	Useless	Severe Criticism

Classify medical problems into three different groups of sentiments:

- Existence-User shares every health problem's symptoms.
- Restore-The user reveals the rehabilitation status from past medical issues
- Depreciate- The user describes his medical condition for deteriorating over the medical treatment period.

The classification strategy focuses on the medication's effect:

- Fruitful-User expresses the positive sentiment in the form of treatment's utility.
- Useless- The user narration reports the therapy's no effect.
- Severe Criticism- User expresses the negative opinion mainly in the form of adverse drug effect against treatment.

In order to obtain potential and effective sources which satisfy the above requirements, we did exhaustive search exploiting multiple medical forums.

In total we collected 1,00000 blog posts including 51,880 posts on medical conditions and 23,020 posts on blog posts related to medication removed 25,100 blog posts where medication or medical condition was not mentioned.

Testing different machine learning strategies (Support Vector Machine (SVM), Naive Bayes (NB)) to determine our system's pre-eminent classifier.

A WEKA, a software platform for data mining techniques and machine learning algorithms that includes several implementations of different models of existing, has been employed to test these techniques. The results obtained with the proposed method are tested in such a chapter to measure the incidence of different diseases as well as to compare the efficiency of the different classifiers. In addition, presented results for the traditional precision, F-measure, Recall and ROC curve.

Precision: The ratio of cases marked by an model as positive that are really positive;

Recall: The ratio of positive instances are properly marked as positive;

F-measure: The harmonious measure of accuracy and recall, i.e. F=2 *(Precision * Recall)/(Precision+ Recall);

Receiver operating characteristic (ROC) curve: At different threshold settings, the ROC curve describes the rate of true and false positives.

Table II Classification of results acquire from Subsets of features generated using Coorelation based feature selection

Classification Strategy	Classification Models		Precision	Recall	F-Measure	ROC
Complication	Sentimental word	NB	0.78	0.75	0.76	0.72
		SVM	0.89	0.8	0.82	0.79
Therapeutic	Sentimental word	NB	0.71	0.69	0.68	0.7
		SVM	0.82	0.78	0.75	0.77

Table III Dataset statistics for section 1 Complication

Section 1: Complication				
Existence	Restore	Depreciate	Average of sentences	Average of words
21960	7100	22820	15	185

Table IV Dataset statistics for section 2Therapeutic

Section 2:Therapeutic				
Fruitful	Useless	Severe Criticism	Average of sentences	Average of words
4520	5130	1330	12	170

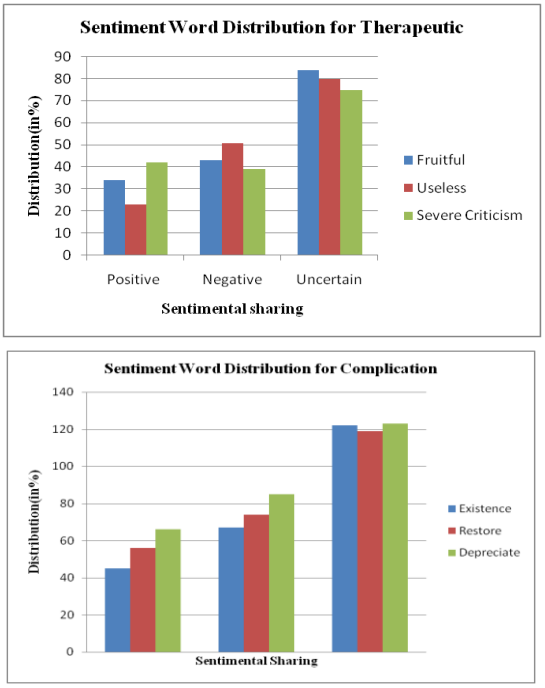

Fig.2. Sentiment Analysis through Sentimental Words

V.CONCLUSION

Health sentiment analysis aims at identifying the primary health facilities and determining what people like about them or hate them. It is a very powerful system to explicitly express the appreciations of patients to those concerned and to inspire them to do well in the future. Sentiment analysis for twitter data shows high precision but low recall value while conducting using Lexicon-based approach, resulting in performance issues. Combining both strategies, i.e., to increase the efficiency. Lexicon based approach and Machine Learning, this gives better results.For the different size training datasets, while conducting the experimental analysis, Such empirical results suggest that the proposed algorithm is highly effective and good for twitter message sentiment analysis.

REFERENCES

1. 1.Schwartz, J.E. and A.A. Stone: 1998, 'Data analysis for EMA studies', Health Psychology 17, pp. 6 –16.
2. Yoon, S., Elhadad, N., and Bakken, S. (2013). A practical approach for content mining of tweets. American journal of Preventive Medicine, 45(1),122-129. doi:10.1016/j.amepre.2013.02.025
3. 3. Nasraoui, O. (2008). Book review: Web data mining- exploring hyperlinks, contents and usage data. SIGKDD Explorations, 10(2), 23-25.
4. Paul, M.J.; Dredze, M. A model for mining public health topics from Twitter. Health **2012**, 11, 16.
5. 5. Denecke, K.; Krieck, M.; Otrusina, L.; Smrz, P.; Dolog, P.; Nejdl, W.; Velasco, E. How to exploit twitter for public health monitoring. Methods Inf. Med. **2013**, 52, 326–339.
6. 6. Behera, P.N.; Eluri, S. Analysis of Public Health Concerns using Two-step Sentiment Classification. Int. J.Eng. Res. Technol. **2015**, 4, 606–610.
7. 7. Missier, P.; Romanovsky, A.; Miu, T.; Pal, A.; Daniilakis, M.; Garcia, A.; da Silva Sousa, L. Tracking dengue epidemics using twitter content classification and topic modelling. In Proceedings of the 16th International Conference on Web Engineering, Lugano, Switzerland, 6–9 June 2016; pp. 80–92.
8. 8. Adrover, C.; Bodnar, T.; Huang, Z.; Telenti, A.; Salathé, M. Identifying adverse effects of HIV drug treatment and associated sentiments using twitter. JMIR Public Health Surveill. **2015**, 1, 7.
9. Schulz, A.; Mencía, E.L.; Dang, T.T.; Schmidt, B. Evaluating multi-label classification of incident- related tweets. In Proceedings of the Making Sense of Microposts (# Microposts 2014), Seoul, Korea, 7–11 April 2014;
10. Al-Amrani, Y., Lazaar, M., and El-Kadiri, K. E.(2018). Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Computer Science, 127, 511-520.
11. Ribeiro, P. L., Weigang, L., and Li, T. (2015). A unified approach for domai-specific tweet sentiment analysis. 18th International Conference onInformation Fusion (pp. 846-853). IEEE.
12. Ji, X., Chun, S. A., and Geller, J. (2013). Monitoring public health concerns using twitter sentiment classification. Proceedingsof the 2013 IEEEInternational Conference on Healthcare Informatics,ICHI'13 (pp. 335-344). Washington, DC, USA: IEEE Computer
13. Salas-Zarate, M. P., Medina-Moreira, J., Lagos-Ortiz,K., Luna-Aveiga, H., Rodriguez-Garcia, M. A., and Valencia-Garcia, R. (2017). Sentiment analysis on tweets about diabetes: An aspect-level approach.Computational and Mathematical Methods inMedicine, 9 pages. doi:10.1155/2017/5140631
14. Myslín, M.; Zhu, S.-H.; Chapman, W.; Conway, M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J. Med. Int. Res. **2013**, 15, 174.
15. Bosley JC, Zhao NW, Hill S, et al. Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication. Resuscitation. 2013;84(2):206–212. doi:10.1016/j.resuscitation.2012.10.017

AUTHORS PROFILE

	**First Author** E.M.Roopa Devi is received the B.E(Electrical and Electronics) degree from Anna University and also M.E(Computer Science and Engineering) degree from Anna University . She has 9 years of experience in teaching field. Her research interest are in the area of Computer Networks, Network Security and Data mining.Currently she is working as an Assistant Professor(Kongu Engineering College) and pursuing the Ph.D degree in Computer Science and Engineering. She has published four papers in international journals.
	**Second Author** R.Raja Devi is received the B.E(Computer Science and Engineering) degree from Bharathiar University and also M.E(Computer Science and Engineering) degree from Annamalai University . She has 15 years of experience in teaching field. His research interest are in the area of web Technology ,Service Oriented Architecture ,Web services ,Optimization and Data Mining.Currently she is working as an Assistant Professor(Kongu Engineering College) and pursuing the Ph.D degree in Computer Science and Engineering. She has Published four papers in international journals.
	**Third Author** S.Vinoth Kumaris received the B.E(Electronics and Communication) degree from Anna University and also M.E(Computer and Communication) degree from Anna University .He has 6 years of experience in teaching field. His research interest are in the area of Computer Networks and Network Security.Currently He is working as an Assistant Professor in Kongu Engineering College.