Indian Journal of Science and Technology, *Vol 9(6), DOI: 10.17485/ijst/2016/v9i6/81691, February 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Advocacy Monitoring of Women and Children Health through Social Data

G. R. Ramya* and P. Bagavathi Sivakumar

Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore - 641112, Tamil Nadu, India; gr_ramya@cb.amrita.edu, pbsk@cb.amrita.edu

Abstract

Background/Objective: To classify, the extracted women and children health data, from the social media, and to utilize it for advocacy monitoring. **Methods/Statistical Analysis:** Advocacy monitoring can be performed by extracting the social data related to women and children health. A keyword based search technique is used for this purpose. The children health details like the nutrition deficiencies, lack of vaccination, diseases like pneumonia, diarrhea and malaria that affect new born children and the women health data like maternal weight loss, maternal mortality rate, sanitation and antenatal care during maternity can be gathered from the social media using keyword based search technique. The extracted data are needed to be analyzed and classified into related data groups using Decision tree C4.5 and Support Vector Machine (SVM). **Findings:** Decision tree C4.5 algorithm classifies the data based on the concept of information entropy. The data are classified at each node of the tree after analyzing the attribute of the data. SVM analyzes the extracted data and uses the health parameters listed to group the related data. The approach is of two stages: training and testing. The training dataset is build using the health data representing the listed search words. This training set is used to classify the test data. The data are tested with the training set and only women and child health data are stored in classes that help in advocacy monitoring in an efficient way. **Applications/Improvements:** Advocacy monitoring is required to define the socio-economic status of a region. The proposed approach efficiently classifies the extracted social data of women and children health and aids in effective advocacy monitoring.

Keywords: Advocacy Monitoring, Decision Tree C4.5, Search Approach, Support Vector Machine

1. Introduction

Advocacy monitoring is an important process in the social and economical development of a country. The monitoring of the social equalities of women and children is very essential to improve the social infrastructure. The social equality also includes the education and health benefits to the women and infant children[1]. Advocacy can be defined as the efficient use of information to influence the policies[2] and actions of the responsible authorities which can produce affirmative changes in the lives of people. The objective of this research is to provide an efficient method to enhance the monitoring of women and infant health with less complexity[3]. In this paper, the women and child health data are extracted from the social media by utilizing a search approach. The approach provides a better aggregation of data as per the search keyword given. Many techniques have been used for efficient social data extraction.

Disease related blogs are often used by people for the betterment of their knowledge. In[4] presented Google Influenza Like Illness (ILI) analysis techniques based on the search queries. The approach provided a strongly correlated pattern for the center for disease control and prediction. The technique uses the English language English language terms which are perfectly matched with the words influenza and flu. The misspelt words are not included. These terms are utilized to gather the blog content related to flu like details discussed by the members or messages posted in the columns. Likewise when the search data is gathered they are analyzed and some words like ill, sick are used to categorize them.

*Author for correspondence

In[5] presented a similar technique called automatic flu extraction by ILI text markers. The technique uses text based extraction in the social media instead of the personal blogs. The web data aggregation is very popular as it provides higher data content for analysis. A new tree matching based approach for the web data extracted has been presented in[6]. The technique collects the data as per the search and analyzes them to arrange in a tree format for easier processing. The tree matching technique is more efficient in extracting the HTML data. The aggregated data includes all types of data from the social media which are relevant to the search but not sufficient for the monitoring for women and children health. Hence a supervised learning model is introduced to mine the data[7]. Thus the approach can provide efficient data that helps in advocacy monitoring.

In [8]presented an automatic automatic approach to monitor the customer satisfaction using the social data. The approach uses filter technique to trace the data related to the customer comments and reviews that not only present in the official website but also in the personal communications about the product or the service. The main focus of the technique would be to extract product review data from the survey contents and message sections. The technique uses survey agent tracker to aggregate the survey contents that has keywords of the product reviewed. The survey analyzer and response analyzer are used to segregate the data that helps in estimating the customer satisfaction level.

In[9] presented an efficient comparison technique to monitor the social inequalities regarding the health of men and women. The technique gathers the data from various sources like government data, print media and social media. These data are analyzed in such a way that the datasets are constructed using the optimal data. The objective of the author is to prioritize the reasons for the inequalities between the two genders of a region. The health inequalities include the differences in the life expectancies and the reasons for the same. The comparison technique presented though has no specific techniques for children health.

Web third-person effect has been the backbone of the web applications including the social media. In[10] presented an approach to prove the effectiveness of third person effect in the media websites. A survey consisting of basic queries linked to the online usage related to the social interactions has been conducted to gather the data. These data are analyzed to determine the role of web third person effect. In[11] and[12] utilized the benefits of opinion mining and opinion analyzer to extract data from websites and micro-blogs respectively. In[13] an efficient machine learning algorithm is used to extract the data from social media messages. These techniques were highly effective in data extraction. By studying various techniques in detail, the work plan to extract and utilize the social data for advocacy monitoring of women and children health can be framed.

2. Methodologies

In this section, the methodologies used in the research for advocacy monitoring are discussed. The pre-processing stage is the removal of the unwanted symbols in the extracted data including the greet words like hai, wow, etc and has tags.

The Decision tree algorithm C4.5 is used to classify the data extracted from the social media based on information entropy. The decision tree is built from the training data created from the already classified samples. The training data is taken as S with samples s_1, $s_2 \ldots s_n$. Each sample of the training data consists of a dimensional vector which contains the values of the instances and the class to which they belong. At each node of the tree, the algorithm selects the attribute of data with high information in one of the classes. This process continues until all the nodes are processed to classify the data. The technique is efficient in the classification of the women and children health data to aid the advocacy monitoring.

The Advocacy monitoring using SVM methodology consists of two levels. In the first level, the data from the social media is extracted using the simple search approach. The search queries are used to collect the data. In the second level, an efficient supervised model called Support Vector Machine is used to classify the raw data into into different classesin the advocacy monitoring system.

Initially the data to be collected is marked by the region of concentration i.e. a state or a country or globally. The first level deals with the approach to extract the data from the social media. The social media considered in this research are Facebook and Twitter. The third party applications like Facebook API and Twitter API are used to extract the data from the social media. The API applications provide the access to the social data of multiple social media users in a legal way. The general information, publically posted comments and other contents that contain the search words are collected by the API.

The extraction process includes the search generation to retrieve the related details. The approach is an automatic extraction process that validates the health data extraction from the gathered data. The approach finds the degree of relatedness between the search word and the gathered data.

Let us consider a survey like keyword list q. The keyword list q contains a set of words that are either directly or indirectly related with the women and children health. It can also be just simple messages like 'maternity death'. The media websites list the search results which can be collected. The collected set contains all the data related to the search. Usually some of the queries are important to gather women and children health data. The search words include nutritional deficiencies, diseases like pneumonia, diarrhea, malaria, are essential to collect data that are relevant to infant children. Likewise weight loss during pregnancy, maternal mortality, sanitation related effects and lack of clinical antenatal care during maternity are essential in collecting women health data. These search words are sufficient enough to collect the women and children health and further analyze to construct the dataset. The data are gathered using q search are both important as well as just relevant terms. The concept of term frequency is utilized to filter the meaningful data. The approach is that the most common content with less relevance to the search words are abandoned before classification.

As the data collected contains both relevant and irrelevant data about women and children health, they are needed to be grouped into classes as per the search words. The normal search based classification cannot be used as the method is not so significant. The data are needed to be classified in terms of relevance to aiding advocacy monitoring. In the second level, the decision tree C4.5 and Support Vector Machine (SVM) is used to classify the data as per the relevance to search. Decision tree C4.5 usually classifies the data at each node. SVM analyzes the health data using training set and testing dataset. A training set is constructed using the standard health data as per the listed search words. Then the raw (test) data are compared with the training data and classified into groups. This data can be utilized for continuous advocacy monitoring.

The advocacy monitoring of women and children health data can be performed using the classified data. The data are compiled into statistical reports and can be utilized for developmental activities. The performances of the two classification techniques are analyzed to evaluate the efficient classification of the given data.

2.1 Algorithm

Advocacy monitoring using SVM.

Input: Unclassified health data from Social media.

Output: Classified women and children health data.

Initialize:

Set of search words $q = \{q_1, q_2, q_3, \ldots\ldots q_n\}$ (1)

Select social media (Facebook, Twitter).
Facebook API and Twitter API.

Gather search results $S_i = \{S_1, S_2, S_3, \ldots\ldots S_n\}$ (2)

Set Q as factor determining the meaningfulness of search results.
For each S_i

If $Q \geq 0.01$ (Q below 0.01 for any search result is meaningless).

Add S_i to the meaningful search set M.
End if.
End for.
Degree of relatedness.

$$Relatedness\ R(q, M_i) = \frac{\sum_{l=1}^{j} v_{q,l} v_{M_{i,l}}}{\sqrt{\sum_{l=1}^{j} \left(v_{q,l}\right)^2} \sqrt{\sum_{l=1}^{j} \left(v_{M_{i,l}}\right)^2}} \quad (3)$$

Where l = 1 to j is the size of the meaningful search set M; $v_{q,l} v_{M_{i,l}}$ are the vector functions of q and M_i.
//Decision tree C4.5//
Assign M to N nodes.
Find attribute with best information gain.

$$Gain\ (X, c) = Info\ (X) - I(c) \quad (4)$$

// where X is the attribute and c is the cut point of the attribute X.
Split M with respect to $Gain\ (X, c)_{best}$. (5)
//SVM//
For each feature satisfies M_i satisfies $g(x) = 0$

$$g(x) = \sum_i M_i \cdot w(x) - b \quad (6)$$

//where w(x) is the vector function of hyper plane; b is the hyper plane.
If $g\ (x) \geq 1$, i = 1, 2… n. (7)
Split M_i into m classes.
End if.
End for.
Compile statistical reports.
Evaluate the performance.

2.2 Description

The set of search words is initially generated. Facebook and Twitter are selected as the source of social data. The API is used to collect the data effectively. The search words are introduced into the social media website to search the related content. The degree of relatedness is computed by comparing the search words and the data gathered. The data with high degree are selected and decision tree C4.5 and SVM are used to classify them into groups. Using the classified data the statistical reports are generated to perform continuous monitoring.

3. Experimental Results

The Facebook API and Twitter API are used to collect the data from the social media with high degree of relatedness. The number of data taken is 200. The results vary with the change in number of data and hence the data below 200 are considered. The experiments are conducted using the data collected. The presented technique is compared with the decision tree algorithm C4.5 in terms of accuracy, precision and recall. The comparison graphs are given below:

3.1 Accuracy

Accuracy of women and children health data classification is the exact classification of relevant data and irrelevant data from the total classified data. Accuracy is evaluated as:

$$Accuracy = \frac{(True\ Positive + True\ negative)}{(True\ positive + True\ negative + False\ positive + False\ negative)} \quad (8)$$

The corresponding result of the presented technique for advocacy monitoring is evaluated. Figure 1 shows the comparison of Advocacy monitoring using SVM and decision tree algorithm C4.5 in terms of accuracy. In x-axis the number of data is taken while accuracy in percentage is taken along y-axis. When the number of data is 25, the accuracy of Advocacy monitoring using SVM is 88% while decision tree algorithm C4.5 is 72%. This proves that the Advocacy monitoring using SVM method has higher accuracy even when the number of nodes increases. This shows that SVM approach can provide effective results for women and children health data.

Table 1 shows the numerical comparisons of the Advocacy monitoring using SVM and decision tree algorithm C4.5 in terms of accuracy (%). The Decision tree

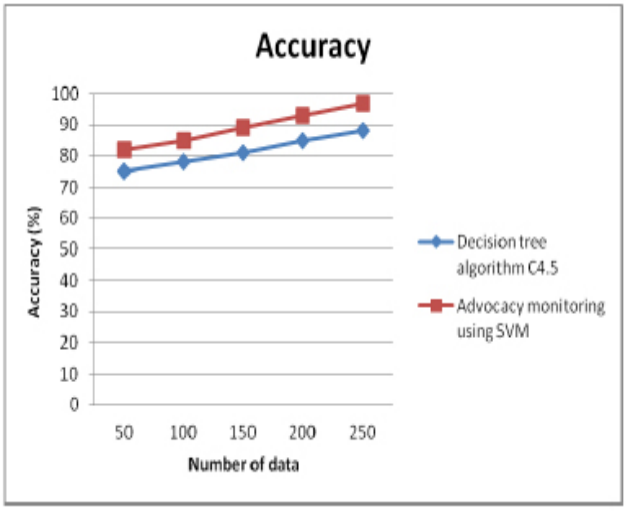

Figure 1. Accuracy.

Table 1. Comparison in terms of Accuracy (%)

Number of data	Decision tree algorithm C4.5	Advocacy monitoring using SVM
50	75	82
100	78	85
150	81	89
200	85	93
250	88	97

C4.5 approach classifies the data into relevant fields with the use of the decision nodes. The SVM based classification approach compares the data with training set and determines the class. From the evaluation results, the classification of social health data with better accuracy is obtained using the SVM approach.

3.2 Precision

Precision value is evaluated according to the retrieval of data.

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \quad (9)$$

The precision of classification for women and children health details through social data can be defined by the ratio of relevant health data to the total calculated relevant data. Figure 2 shows the comparison of Advocacy monitoring using SVM and Decision tree algorithm C4.5 in terms of precision. In x-axis the number of data is taken

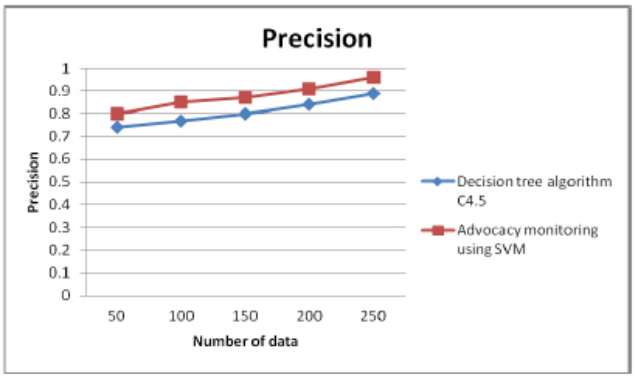

Figure 2. Precision.

Table 2. Comparison in terms of Precision

Number of data	Decision tree algorithm C4.5	Advocacy monitoring using SVM
50	0.74	0.80
100	0.77	0.85
150	0.80	0.87
200	0.84	0.91
250	0.89	0.96

while precision is taken along y-axis. When the number of data is 25, the precision of Advocacy monitoring using SVM is 0.83 while Decision tree algorithm C4.5 is 0.61. This proves that the Advocacy monitoring using SVM method has better precision rate than the state-of-the-art method.

Table 2 shows the numerical comparisons of the Advocacy monitoring using SVM and Decision tree algorithm C4.5 in terms of precision. From the evaluation results, the classification of social health data with better precision is obtained using the SVM approach.

3.3 Recall

Recall value is evaluated according to the retrieval of data.

$$Recall = \frac{True\ Positive}{(True\ positive + False\ negative)} \quad (10)$$

Figure 3 shows the comparison of Advocacy monitoring using SVM and Decision tree algorithm C4.5 in terms of recall. In x-axis the number of data is taken while recall is taken along y-axis. When the number of data is 25, the recall of Advocacy monitoring using SVM is 0.79 while Decision tree algorithm C4.5 is 0.49.

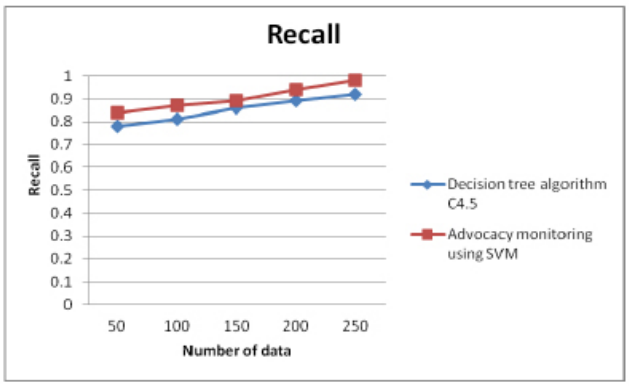

Figure 3. Recall.

Table 3. Comparison in terms of Recall

Number of data	Decision tree algorithm C4.5	Advocacy monitoring using SVM
50	0.78	0.84
100	0.81	0.87
150	0.86	0.89
200	0.89	0.94
250	0.92	0.98

This proves that the Advocacy monitoring using SVM method has better recall rate than the state-of-the-art method.

Table 3 shows the numerical comparisons of the Advocacy monitoring using SVM and Decision tree algorithm C4.5 in terms of recall. From the evaluation results, the classification of social health data with better recall is obtained using the SVM approach.

4. Conclusion

In this paper, the social data is utilized to collect women and children data with less complexity but with high accuracy. In order to perform Advocacy monitoring, a search based approach is used for efficient gathering of women and children health data from the social media (Facebook and Twitter using API). These data are classified using Decision tree C4.5 and SVM for better performance. The data are then utilized to compile statistical report on women and children health in the selected region. Experimental results show that the SVM based technique has efficient performance than the Decision tree algorithm C4.5 in terms of accuracy, precision and recall.

5. References

1. O'Flynn M. Tracking progress in Advocacy: Why and how to monitor and evaluate Advocacy projects and programs. International NGO Training and Research Centre (INTRAC); 2009. p. 1–12.
2. Jayanag B, Vineela K, Vasavi S. A study on feature subsumption for sentiment classification in social networks using natural language processing. International Journal of Computer Applications. 2012 Sep; 53(18):29–33.
3. Westert GP, Schellevis FG, de Bakker DH, Groenewegen PP, Bensing JM, Van der Zee J. Monitoring health inequalities through general practice: The Second Dutch National Survey of General Practice. The European Journal of Public Health. 2005 Feb; 15(1):59–65.
4. Corley CD, Mikler AR, Singh KP, Cook DJ. Monitoring influenza trends through Mining Social Media. BIOCOMP; 2009. p. 1–7.
5. Lampos V, Cristianini N. Tracking the flu pandemic by monitoring the social web. IEEE 2nd International Workshop on Cognitive Information Processing (CIP); Elba. 2010 Jun 14-16. p. 411–6.
6. Kolkur S, Jayamalini K. Web data extraction using tree structure algorithms – A comparison. IJRTE. 2013 Jul; 2(3):35–9.
7. Carvalho JP, Pedro VC, Batista F. Towards intelligent mining of public social networks' influence in society. IFSA Joint World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS); Edmonton, AB. 2013 Jun 24-28. p. 478–83.
8. Erhart G, Matula VC, Skiba D. Method of automatic customer satisfaction monitoring through social media. U.S. Patent Application. 12/777. 2010.
9. Pampalon R, Hamel D, Gamache P. A comparison of individual and area-based socio-economic data for monitoring social inequalities in health. Component of Statistics Canada. 2009 Sep; 20(3):85–94.
10. Antonopoulos N, Veglis A, Gardikiotis A, Kotsakis R, Kalliris G. Web third-person effect in structural aspects of the information on media websites. Computers in Human Behavior. 2015 Mar; 44(1):48–58.
11. Gottipati S, Jiang J. Finding thoughtful comments from social media. COLING; 2012. p. 995–1010.
12. Zhao P, Li X, Wang K. Feature extraction from micro-blogs for comparison of products and services. Web Information Systems Engineering – WISE. Springer, Berlin Heidelberg. 2013; 8180:82–91.
13. Nivedha R, Sairam N. A machine learning based classification for social media messages. Indian Journal of Science and Technology. 2015 Jul; 8(16):1–4.