Indian Journal of Science and Technology, *Vol 8(16), DOI: 10.17485/ijst/2015/v8i16/63640, July 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Machine Learning based Classification for Social Media Messages

R. Nivedha* and N. Sairam

School of Computing, SASTRA University, Thanjavur – 613 401, Tamil Nadu, India;
nivedhagayathri@gmail.com

Abstract

A social media is a mediator for communication among people. It allows user to exchange information in a useful way. Twitter is one of the most popular social networking services, where the user can post and read the tweet messages. The tweet messages are helpful for biomedical, research and health care fields. The data are extracted from the Twitter. The Twitter data cannot classify directly since it has noisy information. This noisy information is removed by preprocessing. The plain text is classified into health and non-health data using CART algorithm. The performance of classification is analyzed using precision, error rate and accuracy. The result is compared with the Naïve Bayesian and the proposed method yields high performance result than the Naïve Bayesian. It performs well with the large data set and it is simple and effective. It yields high classification accuracy and the resulting data could be used for further mining.

Keywords: CART, Classification, Decision Tree, Machine Learning, Twitter

1. Introduction

The social media has now become a center of information exchange. It includes an online communication channel for community-based input, interaction, content sharing and collaboration among people. Social media technologies are in different forms such as internet forum, blogs, picture and music sharing, Twitter, Facebook, etc. The Twitter is a tool for knowledge discovery. The Twitter messages are used in biomedical field in order to find the health and non-health related tweets which can be useful for patients, doctors and researchers. Researchers and general practitioners require social media messages to mine the health related information for various purposes. Data mining is a technique to study data from different perception and combine to form useful information. Machine learning is a technique that can learn automatically and predict the results based on the previous observations. Machine learning based classification provides an accurate prediction and also improves the performance of the results.

2. Previous work

The literature surveys on text classification which is related to our work are discussed below. The social media is the most popular one, which is used to share our thoughts and to communicate with others. Particularly twitter messages are used by the researchers to classify the health related data that can be used for further processing. Collier and Doan[2] collected tweets from the twitter and classify the tweet messages into syndromic categories based on the keywords from the public health ontology. They used the Naive Bayes (NB) and Support Vector Machine (SVM) model to classify the data. It produced a moderate result because the number of training examples is small.

Culotta[4] collected 500 million twitter messages to

*Author for correspondence

identify the flu related messages. The bag-of words classifier is used with logistic regression to predict whether it has a flu related words or not. In that, 95% of the tweets contain the flu related words. Similarly, Corley and Cook[3] did a study on web blogs and identified flu related blog post using the Center of Disease Control (CDC). Yang[10] applied association mining to analyze the relation between the drugs and their reaction. The drugs reaction was identified by detecting the keywords of the Consumer Health Vocabulary (CHV) in social media messages. These methods fail to capture the unknown dictionary words.

Paul and Dredze[6] classified the twitter messages based on the keywords such as disease, symptoms and their treatment using the Ailment Topic Aspect Mode. The SVM was trained with linear kernel, uni-gram, bi-gram, and tri-gram. Joachims[9] suggested that the Support Vector Machine is suited for text categorization after comparing with the K-Nearest Neighbor. Aramaki[1] used Support Vector Machine based classification to identify the flu related tweets and the machine is trained by the uni-gram method. Gharehchopogh[6] did a study on web blogs. The blog is an online diary and it can be easily accessible. The data from the blogs are classified as either professional or non-professional. KNN and Artificial Neural Network are used to classify the particular person's blogs. The Artificial Neural Network yields 90% of the accuracy. It provided better performance result for blogs.

Paul and Dredze[7] did a similar study to classify public health information using modified LDA technique. This classifier produced a precision of 90.04%. Tuarob, Tucker, salathe[8] combined 5 heterogeneous features set with different classification algorithms to classify the health related data from the Facebook and the twitter messages. The heterogeneous feature sets Dictionary based compound features, Topic distribution features and Sentiment features are combined with different base classifications such as NB, SVM and Random forest. They compared the result with N-gram feature set. Eijk[5] did a study on Online Health Communities (OHC) for Parkinson Net. Parkinson Net is a social network for Parkinson disease where health professionals are collaborated. It uses various types of OHC to deliver the patient-centered care.

3. Methodology

Social media plays an important role in day to day life and all the latest news and other information are easy to communicate among the people. A method for classifying the social media data is depicted in Figure 1.

3.1 Data Extraction

The Twitter is a social networking service and it can be easily accessible by all people. A particular user has to create his/her own account and that user can read and post messages in twitter. It allows the user to post messages of 140 characters or less. Data is extracted from the particular Twitter home TimeLine with permission. The data set is unlabeled, noisy, occupy extra space and also degrade the performance. Hence noisy data should be removed from further process.

3.2 Preprocess

Data pre-processing is a process to remove the unwanted data because it is difficult to classify the raw data. The data extracted from the social media is preprocessed as follows,

- The extracted data may have Retweets. Retweet means different persons post the same messages in the twitter. Hence it should be removed.
- A tweet message may contain text, audio and video. Text messages, links of audio and video can be extracted from the Twitter. The audio and video links must be removed since plain text is our focus
- The punctuations, digits, special symbols and user names are eliminated.
- Stemming is very important in text mining. It is a process of reducing inflected words into word stem. Consider an example user, using, use are the words that represent use. Hence, make all words as root.

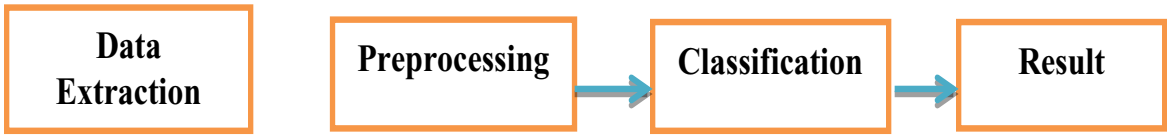

Figure 1. Block diagram.

- The set of words that will not affect the sentences, if it is eliminated is called a stop word. These words should be removed from the text data.
- Case conversion is carried out in order to avoid the multiple ways of represents same words.

Finally, the plain text is extracted from the tweet message. Then the plain text is manually labeled, to simplify the classification process.

3.3 Classification

Classification is a technique used in data mining to classify the data into multiple classes. The classification has two steps, namely training and testing. Training means making the machine to learn with the data and their associate class. Testing means making the machine to test the data based on the previous observation. In classification, the decision tree is most popular and simple technique to classify the data into two or multiple classes. The tree is trained by growing decision trees and tested by following the path of the current tree. It provides an accurate result and easy to handle multi-dimensional data. Classification and Regression Tree is a type of decision tree and a non-parametric learning technique. It allows only binary slit. It will start from the root node and then it search for a split among all the possible values. The recursive partition will take place for a subset of the tree. The tree growing process is repeated until the stopping criteria are met. Then prune the tree in order to reduce over fitting.

3.4 Algorithm

Input:

Data set D.

Gini index used for splitting_critertia.

Output: Decision Tree.

Method:

Determine the splitting_critertia.

H ← φ

NH ← φ // H is the health class and NH is the non-health data class.

For each d € D do

begin

Apply splitting_critertia.

If D is a health data then

H = HU[d]

Else

NH = NHU[d]

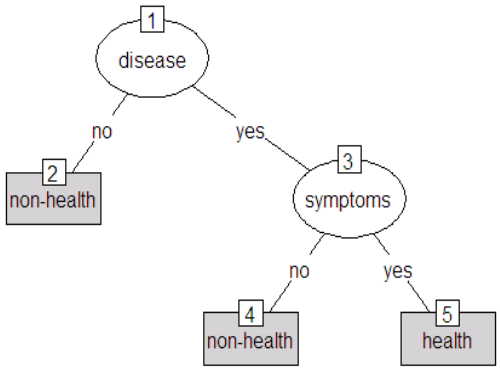

Figure 2. Sample Tree.

This algorithm can be invoked by passing a Data set D and gini index as parameters. The splitting_critertia based on gini index helps to choose the best attribute splits. To create a root node, initially the algorithm will check whether the information contains a disease or not and if it contain the disease then it is classified as health class otherwise non-health class. This process is repeated for each subset of the tree. Figure 2 represents a decision tree for twitter data set and it is implemented using R programming language.

4. Experimental Result and Performance Analysis

Sample tweet messages are extracted from a particular Twitter account after getting prior permission. The extracted data has 1871 words. After preprocessing, the number words are reduced to 1308. It is manually labeled for classification. The labeled data are classified using the CART algorithm. The performance of the classification is analyzed using precision, error rate and accuracy as follows,

Precision = $TP \div (TP + FP)$

Error rate = $(FP + FN) \div (TP + TN + FP + FN)$

Accuracy = $(TP + TN) \div (TP + TN + FP + FN)$

Where, TP, TN, FP and FN denote True Positive, True Negative, False Positive, and False Negative respectively. Table 1 shows the performance of CART and NB. The proposed methodology yields high performance when compared to the Naïve Bayesian classifier. It achieves 84% of accuracy and has an error rate of 15.7894%, which is better than the NB classifier. It acquired high precision

Table 1. Classification Result

MEASURES	CART	NAIVE BAYESIAN
Precision	75.00000 %	35.29412%
Error Rate	15.78947%	22.72727%
Accuracy	84.21053%	77.27273%

when compared to the NB classifier. It obtained a better result for Twitter data.

5. Conclusion

This paper has proposed a decision tree classification on Twitter data to classify the health and non-health related data. The result is compared with the Naïve Bayesian classifier. The proposed methodology gives classification accuracy better than the NB classifier and performs well with the large data set. It is simple and effective. The resultant data are helpful for biomedical and research fields for further mining.

6. References

1. Aramaki E, Maskawa S, Morita M. Twitter catches the flu: detecting influenza epidemics using twitter. Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP'11); 2011; Stroudsburg, PA, USA: Association for Computational Linguistics; p. 1568–76.
2. Collier N, Doan S. Syndromic classification of twitter messages. Electronic Healthcare. 2012; 91:186–95.
3. Corley C, Cook D, Mikler A, Singh K, Arabnia HR. Using web and social media for influenza surveillance. Advances in Experimental Medicine and Biology. 2010; 680:559–64.
4. Culotta A. Detecting influenza outbreaks by analyzing twitter messages. 2010 Jul 27.
5. Eijk MVD, Faber JM, Aarts WJ, Kremer AJ, Munneke M, Bloem RB. Using online health communities to deliver patient-centered care to people with chronic conditions. J Med Internet Res. 2013; 15(6):e115.
6. Gharehchopogh FS, Seyyed SR, Maleki I. A new approach in bloggers clasification with hybrid of k-nearest neighbor and artificial neural network algorithm. Indian Journal of Science and technology. 2015; 8(3):237–46.
7. Paul MJ, Dredze M. A model for mining public health topics from twitter. Technical Report; 2011. p. 1–7.
8. Paul MJ, Dredze M. You are what you tweet: Analyzing Twitter for public health. 5th International AAAI Conference on Weblogs and Social Media; 2011. p. 265–68.
9. Tuarob S, Tucker CS, Salathe M, Ramd N. An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. Journal of Biomedical Informatics. 2014; 49:255–68.
10. Joachims T. Text categorization with Support Vector Machine: learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning (ECM'98). 1998; 1398:137–42.
11. Yang CC, Yang H, Jiang L, Zhang M. Social media mining for drug safety signal detection. Proceedings of the ACM International Workshop on Smart Health and Wellbeing (SHB'12). New York, NY, USA: ACM; 2012. p. 33–40.