Volume 6, Issue 2, February 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Algorithms for Opinion Mining and Sentiment Analysis: An Overview

G. Sneka*
P.G Scholar, Department of Computer Science
Avinashilingam University, Coimbatore,
Tamilnadu, India

CT. Vidhya
Assistant Professor, Department of Computer Science
Avinashilingam University, Coimbatore,
Tamilnadu, India

Abstract*— The emergence of social media and the web has lead to the sheer volume of resources through the various discussion forums, blogs and media websites. The users are thus striving hard to understand and discover huge amount of reviews over the web and thus researches in sentimental analysis is growing rapidly. It becomes impossible to manually analyze the reviews and it became a motivation for effective opinion mining. To overcome such problem there is a need for effective algorithms to be used in the predicting the sentiments that paves the way for important decision making for the users and the manufacturers in the feedback of their product. This paper provides the survey about the challenges and overview of some classification and clustering algorithms used for sentimental analysis and opinion mining.

Keywords*— Opinion Mining, Sentiment analysis, Web mining, Data mining, Text mining.

I. INTRODUCTION

Web mining is an area of sub discipline from text mining which aims in mining the semi structured data in the form of Web content mining, Web structure mining and Web usage mining [1]. Sentimental analysis also known as Opinion mining is used in analyzing the important opinion from the reviews generated by the users. When any decisions are to be made regarding the purchase of new product, software or electronic products the people are very much interested in obtaining the reviews from the various websites, blogs or discussion forums. In such case opinion mining or sentimental analysis is used widely which deals with tracking the mood of the people regarding a particular product or topic.

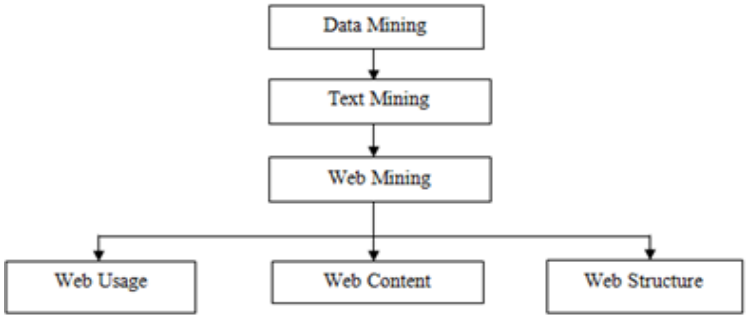

Fig. 1 Data Mining Hierarchy

A topic can be of event, movie, location, drug, product, hotel e.t.c. It is a Natural Language Processing and Information Extraction that aims in obtaining the feelings of the writer that are given by the positive or negative comments, by analyzing a huge volume of documents. While extraction of the data any types of classification, clustering, association rule mining, regression methods can be employed for the particular model chosen. In each method it contains the specific algorithms and according to the type of usage and application area required algorithms can be used.

II. TECHNIQUES IN OPINION MINING

The data mining algorithms can be classified into different types of approaches as Supervised, Unsupervised or Semi - supervised algorithms. Supervised approaches works with set of examples with known labels. In unsupervised approaches aims to obtain the similarity of the attribute values without knowing the labels of the example in the dataset. Semi supervised approaches are being used when the examples in the dataset is the combination of both the labelled and unlabelled examples [4].

A. Classification

Classification is the Supervised technique in which every instances belongs to the specific class, it is being indicated by the value of class attribute or any special goal attribute. The categorical values are taken by the goal attribute in which each attribute belongs to the corresponding class. Two different parts that exist in each example are set of

predictor attribute values and goal attribute value. In classification technique the mining function can be classified into set of tasks such as the training and test set.In the training phase the model that is to be used for the effective classification will be formed up from the training set and in testing phase the model will be evaluated on the test set. Main goal of the classification algorithm is to improve the predictive accuracy in training the model.
The algorithms being discussed includes the following

- K-Nearest Neighbor,
- Support Vector Machines,

1) K-Nearest Neighbor: K-Nearest Neighbor algorithm that is being widely used for classification and regression and also it is a non-parametric method. Every training set that is being present in the multidimensional feature space are the vectors with the specific class labels specified. In n-dimensional space each attribute will be pointing to the training samples with n dimensional numeric attributes. The training phase of the algorithm it acts by storing the feature vectors and class labels. When an unknown sample is given to k-nearest neighbor algorithm it searches for the pattern space for the k training samples that are closer to the unknown samples [7].Euclidean distance determines the property of the "closeness" measures.When KNN approach is to be applied value should be appropriate and the effectiveness of this approach mainly depends upon this value [2] [3].

Advantages of KNN Algorithm

- It can be widely adopted for multi-class model classes and also for the objects with multiple class labels.
- KNN is an efficient classification algorithm that is easy for understanding.
- Building of the model is also inexpensive with extremely flexible classification.
- It is robust even in the case of large dataset with noisy training data being used.

Application of KNN Algorithm

It is widely used in the application areas of legal, medical, agriculture, news and banking for problem solving, functional learning and for user training purposes(Teaching and aiding).

Finance:

- KNN plays an important core role in identifying the stock market forecasting such as analysing the market trends, planning for investment strategies and identifying the time period for obtaining the stocks and for credit ratings.
- Used in the banks for loan management, bank customer profiling.

Medicine:

- To predict patient's health based on the clinical record persisting to each patient. For example, in estimating the amount of glucose level in the diabetic patients from the existing health record.
- When coupled with genetic algorithms, KNN is being used for analyzing the micro-array gene expression data.

Agriculture:

- Used in the climate forecasting and to estimate the soil water parameters.
- In simulation and precipitation of weather variables the KNN algorithm is widely used.

2) Support Vector Machines: SVM was introduced by Boser, Guyon and Vapnik and widely being used for classification, regression and pattern recognition.SVM has capability to classify indeed of the dimensions or size of the input space. It acquires the major advantage because of its high generalization performance with indeed of the much prior knowledge. The goal of the SVM lies in finding the best classification function and also it aims to distinguish between members of the two classes in training data. The major idea behind the SVM is construction of the optimal hyper plane that is widely used for the problems of classification and for patterns identification .From the set of hyper planes the hyper plane that is of optimal is needed to be selected for pattern classification and thus to improve the margin of the hyper plane. SVM needs to classify the given patterns correctly so that it can maximize the margin that determines the efficiency of the SVM algorithm. The accuracy in classifying pattern will improve based on the size of the margin i.e. greater the margin size more exactly it classifies the patterns [7]. The equation for the hyper plane is given below.
Hyper plane, $aX + bY = C$,
With the help of kernel function $\Phi(x)$.i.e. $\longrightarrow \Phi(x)$
,the above pattern can be mapped into high dimensional space.SVM tries in finding the hyper plane accurately that separates the two different samples with the set of independent training samples being specified [5].

Advantages of SVM algorithm

- It provides the greater benefits on the text classification when the high-dimensional spaces are being used.
- Accuracy in the prediction is comparatively high with other classification algorithms.
- Fast evaluation of the learned target function.
- Used widely in various real time applications with the high scope in evaluating the good outcome.
- Without the dependence of the dimensionality of feature space it has the good ability in learning.
- It interprets the inherent characteristics of the data better when comparing to artificial neural networks.

Application of SVM algorithm

It is been widely used for many real world problems such as

Text categorization:

- To categorize the text documents i.e. natural text, based on their content, for example in email filtering, web searching, sorting the documents to specific topic.
- In assigning documents to more than one category such that for series of binary classification problem.

Image classification:

- Used in validating and testing the bacterial image, pathogens and for the classification SVM is used widely.

Medicine:

- It is used in detecting the micro calcifications in mammograms which is an indicator for the breast cancer.

B. Clustering

Clustering is the unsupervised technique that performs natural grouping of instances. It is the method of dividing the data into different groups with the similar objects. Every group, called cluster, consisting of several objects that are similar within the particular cluster and dissimilar to the objects of the other clusters. Clustering algorithms are also used for data compression too rather than the categorizing and organizing the data. An effective clustering algorithms aims in obtaining the effective clusters irrespective of their shapes and size of data. Most Commonly used algorithms in the clustering falls into any of the following categories as Hierarchical, Partitioning, Grid based, Density based, Model Based and Constraint based algorithms.

The algorithms that are being described in this paper are

- K-Means Clustering
- SOM(self organized map)

1) K-means clustering algorithm: K-means algorithm is most common and popular clustering tool that is widely used in many applications and it falls under the partitioning algorithms that aims in constructing the various patterns and evaluates them by using some criterion. With the given collection of n data, k different clusters are formed with each cluster having an unique centroid (mean) and thus the partitioning is made. The letter k describes the number of clusters needed to be made. When number of n objects is to be grouped into k clusters, K cluster centre is to be initialized. Every object will be given to the closest cluster centers and .the centre of cluster is updated every time until state of no change occurs in the each cluster. The elements in each cluster will be in close contact with centroid of that particular cluster and will be different to the elements belonging to other clusters [6].

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} | p - m_i |^2$$

The sum of the discrepancies between the point and the centroid expressed by specific distance is used as the objective function. Total intra-cluster variance describes the sum of the squares of the error between the point and respective centroids.

Advantages of K-means clustering algorithm

- The k-means algorithm produces an relatively scalable results when handing the large data sets.
- It is used for undirected knowledge discovery and very simple to use.
- Used widely in various number of applications ranging from unsupervised learning of neural networks, image processing, pattern recognition and may other applications.
- It can provide the best results when the datasets are distinct and well separated from each other.

Applications of K-means clustering algorithm

- Used in acoustic data for converting the waveforms into some category for speech understanding
- Colour based image segmentation is possible by the use of K-means clustering technique.
- It can be used in machine learning and also for the data mining.

2) Self organized Map (SOM) algorithm: SOM is a type of the artificial neural network (ANN) that is unsupervised learning methodology introduced by the professor Kohenon in 1980 s and though it is also called as **Kohonen's Self-Organizing Map**. It is widely used in vector quantization and it belongs to the category of competitive learning networks. The SOM can be used to detect the features that is inherent to the problem and thus it can also been known as SOFM, i.e. Self-Organizing Feature Map. It is majorly used in representing the data in low dimensional discretized representation from the high dimensional representation [8].

SOM consists of numerous components called as nodes or neurons. Each neuron will be assigned with the specific weight in the output space. Based on the weights it will reflect on the cluster content. It provides the topology preserving map from the high dimensional space to map units that are used is preserving the relative distance between the points. In the SOM the points that are near to each other in the input space are mapped to nearest map units-Matrix representation is used in Self organized maps in order to specify or identify the distance between the existing other neurons. With the U-matrix representation the cluster boundries are easily recognized.

Advantages of SOM algorithm

- Since it is the unsupervised learning method it does not need any human intervention but only needs some knowledge about the input data.
- Ability of the network to generalize and characterize the inputs that are not being encountered before.
- It can be applied to compare the various maps of different sizes and can be used effectively in vector quantization.

Applications of SOM algorithm

Used in many real world problem such as

Speech recognition and analysis:

- Used in creating the representation of the spectra for the different speech samples to different parts of the map.
- Visualization property of the SOM is used in the voice analysis applications.

Interpretation of ECG data:

- With the two-dimensional display it is used to monitor the ECG signal as a trajectory.
- By the use of clustering the data decisive features in the sleep ECG is being identified with the SOM.

III. CHALLENGES IN OPINION MINING

- Apart from the noun words, Adjectives and verbs are also considered as feature words in some cases and it becomes difficult to classify [10].
- Customer can use abbreviations, short words or roman letters. For example lex for lexical, cam for camera etc, so it takes time to deal with such type of words and understanding it for the mining process.
- The strength of the opinion various will vary according to the usage area. For The user opinions about various products, feedback would be on different language (French, Chinese, and Greek), so it becomes difficult to tackle each language with its orientation is difficult task and challenging.
- To describe the similar feature, different synonym words will be used and thus it becomes the challenging task [9].
- In sentence level opinion mining it is difficult to classify the sentence as positive, negative or neutral because different people will have various types of writing styles.
- Web contains the various spam contents and it becomes difficult in eliminating the spam and fake reviews before processing to obtain the better accuracy in results.
- example in blogs or discussion forums the strength of the opinion will vary according to the arguments being occurred.

IV. CONCLUSION

Data mining provides the various techniques in identifying the hidden patterns within the huge set of data that can be used in identifying the future behaviours. For obtaining the solution to any type of problems, dataset becomes the key factor and once dataset is choosen any kind of mining algorithms can be explored. Based on type of dataset and the application being used supervised or unsupervised approaches can be used, sometimes the combination of both classification and clustering algorithms can pave the way for better accuracy in results. In this paper we presented some of algorithms that are most widely used in sentimental analysis and opinion mining such as Support vector machine-nearest neighbour, K-means clustering and SOM (Artificial Neural network).

REFERENCES

[1] G.Angulakshmi1 and Dr.R.ManickaChezian,*"An Analysis on Opinion Mining: Techniques and Tools",* International Journal of Advanced Research in Computer and Communication Engineering,Vol. 3, Issue 7,pp .2278-1021, July 2014.

[2] Raj Kumar and Dr. Rajesh Verma,"*Classification Algorithms for Data Mining: A Survey"*, International Journal of Innovations in Engineering and Technology (IJIET), Vol. 1, Issue 2,pp. 2319 – 1058, August 2012.

[3] DelveenLuqmanAbd AL-Nabi and ShereenShukri Ahmed *"Survey on Classification Algorithms for Data Mining :(Comparison and Evaluation)"*, Computer Engineering and Intelligent Systems, Vol.4, No.8, pp 2222-2863,2013.

[4] Poobana S and Sashi Rekha k, *"Opinion Mining From Text Reviews Using Machine Learning Algorithm"*, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 3, pp 2320-9801,March 2015.

[5] Shinaa and NavpreetRupal, *"Review of Classifiers for Automated Opinion Mining"*,International Journal of Computer Applications (0975 – 8887) Volume 97– No.5, July 2014 .

[6] Aastha Joshi and Rajneet Kaur *" A Review: Comparative Study of Various Clustering Techniques in Data Mining"*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013

[7] S.Neelamegam and Dr.E.Ramaraj, *"Classification algorithm in Data mining: An Overview"*,International Journal of P2P Network Trends and Technology (IJPTT) – Volume 4 ,Issue 8,pp 2249-2615,Sep 2013.

[8] AltugAkay.AndreiDragomir and Bj¨OrnErikErlandsson, "*Network- Based Modeling And Intelligent Datamining Of Social Media For Improving Care"*, IEEE Journal Of Biomedical And Health Informatics, Vol. 19, No. 1, January 2015.

[9] Duyu Tang, Bing Qin, Furu Wei, Li Dong, Ting Liu, and Ming Zhou, *"A Joint Segmentation And Classification Framework For Sentence Level Sentiment Classification"*, IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 23, No. 11, November 2015.

[10] Kang Liu, Liheng Xu, and Jun Zhao*,"Co-Extracting Opinion Targets And Opinion Words From Online Reviews Based On The Word alignment Model"*, IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No.3,Mach 2015.

[11] G.Vinodhini and RM.Chandrasekaran ,*" Sentiment Analysis annnd Opinion Mining: A Survey"*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.

[12] Bhuvana,Dr.C.Yamini, *"Survey On Classification Algorithms For Data mining: (Comparison And Evaluation)"*, International Journal of Advanced Research in Science and Engineering,Vol. No.4,Special issue01),August 2015.

