International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

A Study on Data Mining Techniques in Social Media

Shailja Bhardwaj

Department of computer science Graphic Era Hill University, Bhimtal, Uttarakhand, India.

Abstract-Data mining is the procedure of examining pre-existing databases to generate new useful information by using some strategies and implementing particular operations. Accessing social network sites such as Twitter, Facebook, LinkedIn and Google+ through the internet has become very useful and daily part of our life. People are becoming more interested in and depending on social network for information, news and opinion of other users on diverse subject matters [1]. Social Media mining is the process of representing, analysing and fetching actionable patterns from raw data in social media by using many techniques to conquer the problems in social media.

INTRODUCTION

Data mining is the collection of techniques for efficient discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. It is usually used by business intelligence organizations, and financial analysts, but it is increasingly used in the science to get information from the enormous data sets generated by modern experimental and observational methods.

Social Data Mining

Social data mining introduces basic concepts and principle algorithm suitable for exploring enormous social media data, it discusses theories and methodologies from different direction such as computer science, machine learning, social network analysis, data mining, optimization and mathematics. Social media asserts to the growth of these social networks in which individual collaborate with one another through friendship, emails, BlogSpot and many other mechanisms. Social media mining intents to make sense of these individuals enclosed in networks.

Various Techniques of Data mining in Social Media

There are many different types of techniques have been developed to overcome the problems such as size, noise, and dynamic nature of social media data. Due to different types of data and massive volume of data in the social media, it requires an automatic data processing in order to analyse it within a given time span [6]. Different types of data mining techniques are as follows.

I. Unsupervised classification

We can easily decide a review as 'thumbs-up' or 'thumbs-down' by using unsupervised learning. This type of marking can be done locating the phrases including an adjective or adverb. We can estimates the semantic orientation of every phrase by using PMI-IR followed by the grouping of the review by using the mean semantic orientation of the phrase [1].

1.1 Sentiment lexicon

Sentiment lexicon is a collection of sentimental words that are used by reviewers in their expressions. Sentiment lexicon is a catalogue of the common words that intensify data mining techniques. Different aggregation of sentiment lexicon can be created for assortment of subject matters. For example sentimental words used in politics are often different those used in sports. Expanding the occurrence of sentiment lexicon helps to focus more on analysing topic-specific occurrence, but with the use of high manpower, Lexicon-based approaches require parsing to work on simple, comparative, compound, conditional sentences and questions [1].

1.2 Sentiment orientation

Sentiment orientation can be positive, negative, or neutral (no opinion).It might be immense for the future buyers to make the decision regarding the purchase of a product by tracking usable reviews which are attracted by the widespread products. Semantic orientation is also used by the application developers for their application ranking so that they could see the reviews presented by the users. Where the rating is represented in the form of 5-star scale with 5 showing the best ranked while one denotes the poor ranking [1].

1.3 Opinion definition and summarization

These are the important techniques granting opening. Opinion definition can be discovered in a text, sentence or the document's topic, and it can also occupy the whole document. Opinion extraction is difficult for summarization and tracking of any document. Using this technique, the biased (fixed views) part is explored in the texts, and documents [1]. It is required to aggregate the opinion since all the opinions fetched in the document are not as a direct result of consequence concerning the topic under analysis. It plays a vital role in the business organizations and government offices by helping in improving the products and policies respectively [6].

1.4 Basic clustering technique

Clustering can be considered the most important unsupervised learning problem; it deals with finding a structure in a collection of unlabelled data [9]. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Clustering techniques can be applied in many fields, for instance: Marketing, Biology, Libraries, Insurance, City-planning, Earthquake studies, and WWW (World Wide Web). Clustering techniques involves four most used clustering algorithms; K-means, Fuzzy C-means, Hierarchical clustering, Mixture of Gaussians. So that, K-means is an exclusive clustering algorithm, Fuzzy C-means is an overlapping clustering algorithm, Hierarchical clustering is obvious and lastly, Mixture of Gaussian is a probabilistic clustering algorithm [12].

1.5 Opinion extraction

This technique is compulsory in order to aim that chunk of the document including genuine opinion. An individual's opinion regarding a skilled subject does not matter unless that particular individual has mastered that specific domain [6]. However, the use of both opinion extraction and summarization is essential because of the opinion from many people. The massive number of people giving their opinion regarding a certain subject, it will be more significant to take out that particular [1].

Other types of unsupervised learning which are being used nowadays are POS (Parts of Speech) tagging. Sentiment polarity is the binary classification technique that classifies the opinionated document into predominantly positive or negative opinion (Adedoyin-Olowe, Gaber, & Stahl, 2013).

2. Semi-supervised classification

Semi-supervised classification focuses on enhancing supervised classification by minimizing errors in the labelled examples, but it must also be compatible with the input distribution of unlabelled instances.

Semi-supervised classification can be categorised into two slightly different approaches, Transductive and Inductive learning. Transductive learning concerns the problem of predicting the labels of the unlabelled examples, given in advance, by taking both labelled and unlabelled data together into an account to train a classifier. However Inductive learning considers the given labelled and unlabelled data set as the training examples, and its objective is to predict unseen data [13].

3. Supervised classification

A co-occurrence of collectivegrounds of piece of information is used by supervised learning algorithm in order to mark many adjectives distinguished by alike or unlike semantic orientations [2]. Supervised classification can be very effective and accurate in classifying satellite images and can be applied at the individual pixel level or to image objects (M.A & Alhamad, 2006).

3.1 Support vector machine

Support vector machines are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. It can be used to solve various real world problems such as text and hypertext categorization, hand-written character recognition, and classification of images [4].

3.2 Neural Network

Neural networks are well-suited to identifying non-linear patterns, as in patterns where there is not a direct, one-to-one relationship between input and the output, neural network consists three layers for work to achieve its goal, and these layers are Input Layer, Output Layer, and hidden Layer. Neural network is basically used for predicting pecuniary performance and taking decisions about business [9].

3.3 Naive Bayes

Naive Bayes approach is based on the probabilistic classifiers based on applying strong independent assumption between the features. It can also use in weather fore casting by considering frequencies and cooperation values in historical data. It is highly recommended approach in analysis of sentiments [11].

3.4 K-nearest Neighbor

K nearest neighbors is an easy algorithm which consists of entire training dataset. KNN can be used for both classification and regression predictive problems [6]. However, it is more widely used in classification problems in the industry. KNN approach is also classified into two categories; non-parametric and instance based learning algorithms. Non-parametric means it makes no explicit assumption about the functional form, avoiding the dangers of mis modeling the underlying distribution of the data. Instance-based learning means that our algorithm doesn't explicitly learn a model. Instead, it chooses to memorize the training instance which are subsequently used as "knowledge" for the prediction phase [8].

3.5 Decision tree

One neural architecture is that it is difficult to determine what exactly is going on in the machine learning algorithm that makes a classifier decide how to classify inputs. However, decision trees can present us with a graphical representation of how the classifier reaches its decision. Classification and regression trees framework is responsible to create decision trees [7].

3.6 CHAID (Chi-square Automatic Interaction Detector)

CHAID analysis is an algorithm used for discovering relationship between a categorical response variable and other categorical predictor variables. It is useful when looking for patterns in datasets with lots of categorical variables and is a convenient way of summarising the data as the relationships can be easily visualised. CHAID is often used in direct marketing to understand how different groups of customers might respond to a campaign based on their characteristics [3].

3.7 Text mining

Text mining is the process of extracting useful information from text which will be useful in future scope. Text mining involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging, and information extraction. Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases, but text mining can also work with unstructured or semi-structured data sets such as emails, text documents and HTML files etc. As a result text mining is far better solution [5].

REFERENCES

[1] Adedoyin-Olowe, Gaber, & Stahl, 2013.

[2] Alrababah, M.A., and M.N. Alhamad. 2006. International Journal of Remote Sensing 27: 2703–2718 - used unsupervised and supervised classification methods to map land use, and showed that supervised classification improved map accuracy.

[3] CHAID blog by Sarah Marley, https://select-statistics.co.uk/blog/chaid-chi-square-automatic-interaction-detector/

[4] Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks". Machine Learning.

[5] Daniel Waegel. The Development of Text-Mining Tools and Algorithms, Ursinus College, 2006,http://en.wikipedia.org/wiki/Text_analytics/

[6] Data mining in social media, Thabit Zatari, International journal of Scientific and Engineering Research, Volume 6, 2015

[7]https://pythonmachinelearning.pro/supervised-learning-using-decision-trees-to-classify-data/Decision tree concept.

[8] K-nearest neigbor :https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/