St. Joseph's Journal of
Humanities and Science
ISSN: 2347 - 5331

http://sjctnc.edu.in/6107-2/

ISSN: 2347-5331

St. Joseph's Journal of Humanities and Science

Volume 3 Issue 1 January 2016

Data Mining Techniques Used in Machine Learning

A. Nirmala[a*]
S. Arivalagan[b]
J. Robert Adaikala Raj[c]

ABSTRACT

Today, social media allows a space for the citizen to register their emotional opinion in various sites. Here the opinion is sentiment analysis of users, emotional feelings in micro blogs, reviews, forums etc. Opinion mining is a larger domain for marketing and advertising. Advertiser provides false Star rating information to their product to attract the users. Here Sentiment classification is a specific work of text classification which classify a text according to its sentimental polarities of opinions such as favourable or unfavourable, positive or negative and neutral. Fraud comments could be removed by using a variety of data mining algorithm in Machine learning. At the same instance, concern on users, analyze the information and predicts the information whether it is false or fact in this paper. Machine learning and Lexicon based Classifier uses different classifiers to mine the sentiments.

Keyword: Opinion, Sentiment, Data mining, Machine learning, Lexicon classifier.

1. INTRODUCTION

The sentiment contents posted on social media's like twitter, facebook attracts most of the people to share their opinions and view on various topics. The posted information's on social media can be positive, negative or neutral. Opinion mining is a process to track the view of the public about specific product and their services. An opinion is optimist or pessimist or neutral sentiment, review, view, attitude, emotion, or appraisal about an entity.[1] An entity is an event, organization, product, person or topic. Automated opinion mining implements Lexicon and machine learning. Machine learning is a branch of Natural Language Processing and artificial intelligence (AI) to supervise text for sentiment. Sentiment analysis is done in three levels.

Document Level: The document level is clarified, whether a entire opinion document expressed a positive or negative or neutral sentiment. Here each document is been analysed and it expresses opinions on a single entity (e.g., a single product). Thus, a document level is not relevant to estimate or compare multiple entities.

Sentence Level: Analyzing the sentence in this level to concur on whether each sentence expressed a positive or negative opinion. Here neutral stands for no opinion. The subjective expressions are expressed in various forms e.g. opinions, desires, beliefs, allegations, suspicions, speculations etc.

[a]Research Scholar, Department of Computer Science, Annamalai University, Chidambaram, Tamil Nadu, India.
[b]Department of Computer Science and Engineering, FEAT, Annamalai University, Chidambaram, Tamil Nadu, India.
[c]PG Research Department of Computer Science, St. Joseph's College of Arts & Science (Autonomous), Cuddalore - 607 001.
[*]E-mail: sjcnirmala@gmail.com, Mobile: +91-8825435277.

Text Level: Analyzing the text in this level is to conclude whether each text expressed a positive, negative or neutral sentiment.

An opinion is a fact and it contains subjective information, which may be positive or negative opinion. Objective sentences are emotional sentiment of users. Eg. anger, happy, sad, feeling etc [2]. Others opinions are useful, to make decision on purchasing or advertising the product or guide us to take decision. But the sentiment emotions that are expressed on social media could be false or fact is been analyzed in machine learning by using various algorithm technique in data mining.

The aim of this paper is to assist researchers to expand a better perceptive of data mining techniques used in machine learning approaches and help them in the selection of the right data mining techniques for their research.

2. A LITERATURE REVIEW ON DATA MINING TECHNIQUES IN OPINION MINING

The data mining algorithms can be classified into Supervised, Unsupervised or Semi supervised algorithms. Supervised approaches works with set of examples with known labelled data sets. Unsupervised approaches aim to obtain the resemblance of the attribute values with unknown attribute example unlabeled data set. Semi supervised approach is used when the examples in the dataset is the combination of both the labelled and unlabeled examples.

2.1 Supervised Machine Learning

The Supervised learning works under the supervisor of a set of an input variables (X) and an output variable (Y). And the supervised learning use a CART algorithm to produce an output (Y) from an input (X). (ie) Y = f(X). The main aim of supervised learning is to predict an output variables (Y) from the new input data (x).

CART ALGORITHM

CART Algorithm is grouped into Regression and Classification problems.

Classification: A Classification problem is when an output variable occurs without any overlap in a category, such as "green" or "white" or "happy" and "sad".

Regression: A Regression problem is when an output variable is a real value with overlap, such as "pounds" or "weight"[3].

2.2 Unsupervised Machine Learning

Unsupervised learning has input data (X) and no corresponding output variables. The aim of unsupervised learning is to model the basic structure or distribution in the data, in order to learn more about the data.

Unlike supervised learning, the unsupervised learning has no teacher to produce the correct output data. Unsupervised learning problems can be classified into Clustering and Association problems.

Clustering: A Clustering problem is grouped as type of words and inherent groupings in the data, such as grouping customers by purchasing attitudes.

Association: An Association is a rule-based learning problem, in which we discover the interesting relationships of variables in large portions of data, such as people who buy FISH also tend to buy MEAT[4].

The popular unsupervised learning algorithms are:

- KNN- for clustering problems.
- Apriori algorithm - for association rule learning problems.

2.2.1 K-Nearest Neighbour:

K-Nearest Neighbour algorithm is used for classification and regression . Every training set has multidimensional attribute space for the vectors, specified with the definite class labels. With n-dimensional numeric attributes stored by the feature vectors and class labels with n-dimensional space, where each attribute will be pointing to the training samples.

Unknown sample is specified to search for the pattern space by the k-nearest neighbour algorithm and also to find out the k training samples that are nearer to the unknown samples based on clustering. Advantages of KNN Algorithm are efficient and inexpensive. implemented in multi-class model classes and also for the objects with multiple class labels.

2.2.2 Apriori Algorithm:

The Apriori algorithm uses a associations rules on a set of items in a database to generate the frequent items set and pruned the infrequent item set based on the user opinion.

- A set of all items in a shop are listed as I={I1,I2,I3………..In}
- A set of all transaction database are T={T1,T2………………Tn}
- Each transaction Ti has a set of items
- Each transaction Ti has a transaction ID /TID

A subset of a frequent item set also be a frequent item set

(Eg) {I1, I2, I3, I4 ……….. In} ---------→ is a frequent item set.

{I1}, {I2}, {I3}, {I4}…{In} ---------→ is a frequent item subset.

The Apriori algorithm implements the interactive approach to generate the frequent items listed by the user sentiments on N number of items sets. The algorithm used the iterative approach on large database and extends the N number of items sets into N+1 number of items sets [5].

The advantage of Apriori algorithm is that a subset of a frequent item set can also be a frequent item set. To pruned the infrequent item set.

The Apriori algorithm implements the iterative approach to generate the frequent items listed by the user sentiments on N number of items sets.

The algorithm use the iterative approach on large database and pruned the N number of items sets into N-1 number of items sets.

The Apriori algorithm mine the opinion of user frequent item sets from shop on association rules.

Step 1:

A set of all items in a shop are listed as I={1,2,3,4}. Each transaction Ti has a set of items. Each transaction has a transaction ID /TID={101,102,103,104}

Transaction Database Candidate Items

TID	Items
101	1,2,4
102	2,3
103	1,2,3
104	3

Frequent Items

Items	Support	Items	Support
{1}	2	{1}	3
{2}	3	{2}	3
{3}	3	{3}	4
{4}	1		

The minimum support item is I4=1,

Step 2:

Candidate items Frequent items

Itemset	Support	Items	Support
{1,2}	2	{1,2}	2
{1,3}	1	{2,3}	2
{2,3}	2		

Step 3:

If any sub items are not present in sub-set items, the entire sub items are pruned.A subset belongs to the frequent item set.

Here, {1,2}=>{1},{2}, {2,3}=>{2},{3}, so the {2} and {3} has maximum frequent data solved by associations rule [6].

3. SEMI-SUPERVISED MACHINE LEARNING

In huge amount of input data (X) , only some of the data is labelled (Y) and most of the data are unlabeled between supervised and unsupervised learning, which are represented as Semi-Supervised Machine Learning. The accepted semi-supervised learning algorithm is Lexicon based approach.

3.1 Lexicon-Based Approach

The Lexicon Based Approach deals with expressions, textual content, emotional feelings that occurs in chat, dialogue, discussion, blogs, review, post etc.

The lexicon based approach compares the positive and negative expression to counterpart the words in dictionary to decide the polarity. The Lexicon has classified into Dictionary based approach and Corpus based approach [7].

3.1.1 Dictionary-Based Approach:

This is a effortless technique to processing and analyzing the expression in the dictionary, the positive words are searched in synonyms and the negative words are searched in antonyms[8]. Dictionary based approach searches the Word in Net or a further online dictionary to find out the synonyms and antonyms. The repeated process comes to an end, When we received the message, no more word is here. Then manually it clears the list.

3.1.2 Corpus-Based Approach:

The Corpus-based approach worn to discover a theory (or) hypothesis that aims to draw the text, phrases and opinion in a expert form.

4. CONCLUSIONS

The most important thing to be noticed is the different data mining algorithm implemented in machine learning to analyze the emotional sentiments in wide platform of social media. To reach the enhanced result in machine learning algorithm. still we need to analyze the optimistic and pessimistic result in documents and sentence level by using probability.

REFERENCES

[1] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," J. Am. Soc. Inform. Sci. Technol., vol. 60 (2009), pp. 2169–2188.

[2] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Micro-blogging as online word of mouth branding," in Proc. Extended Abstr. Human Factors Comput. Syst., 2009, pp. 3859–3864.

[3] S.MuthuVisalatchi et al, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.3(2016), pp. 101-107.

[4] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," J. Comput. Sci., vol. 2, Issue 1(2011), pp. 1– 8.

[5] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in Proc. 4th Int. AAAI Conf. Weblogs Soc. Media, vol. 10(2010), pp. 178–185.

[6] L. T. Nguyen, P. Wu, W. Chan, W. Peng, and Y. Zhang, "Predicting collective sentiment dynamics from timeseries social media," in Proc. 1st Int. Workshop Issues Sentiment Discovery Opinion Mining, 2012, p. 6.

[7] M. Thelwall, K. Buckley, and G. Paltoglou, "sentiment in twitter events," J. Am. Soc. Inform. Sci. Technol., vol. 62, no. 2(2011), pp. 406–418.

[8] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in Proc. Workshop Lang. Soc. Media, 2011, pp. 30–38.