International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER
SSRN

IJIEMR Transactions, online available on 06th Jan 2023. Link
:http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=ISSUE-1

DOI: 10.48047/IJIEMR/V12/I1/30

Title **A Survey on Unsupervised, Semi Supervised, Supervised Machine Learning Based Sentiment Analysis**

Volume 12, Issue 1, Pages: 317-326

Paper Authors
ZABI UR RAHAMAN K, Dr. M. GIRI

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

A Survey on Unsupervised, Semi Supervised, Supervised Machine Learning Based Sentiment Analysis

ZABI UR RAHAMAN K[1], Dr. M. GIRI[2]
[1]Research scholar, Department of Computer Science, BEST Innovation University, Gorantla, Sri Sathyasai District, Andhra Pradesh, rahman.naju@gmail.com
[2]Associate Professor, Department of CSE, Joginpally B.R. Engineering College, Moinabad, Hyderabad, Telangana, dr.m.giri.cse@gmail.com

Abstract
Data clustering is used in many applications to classify data samples. Sentiment analysis is used to extract various opinions of customers on different items. Sentiment analysis is one of the trending topics, where both machine learning and artificial intelligence methods are combined used, and it is used in many recommendation applications. The data which is used in sentiment analysis are also plays important role, all these data are collected from customers in the form of reviews, all these reviews are analyzed, and based on results business people take decisions. In this paper, we conducted survey on various unsupervised, semi supervised, supervised based machine learning methods used in the field of sentiment analysis. This study also listed advantages, limitations, and problems faced by machine learning algorithms in sentiment analysis.

Keywords: *SVM, Learning Methods, Sentiment Analysis, Machine Learning, AINN, EA, Clustering*

Introduction

Sentiment analysis is a task of calculating customer's judgments, rating, emotions, sentiments, opinions, topics of interest, and reviews of a particular thing. The other name of sentiment analysis is called opinion mining. Extracting sentiments from online web applications generated dataset is a difficult task. For example, customer rating and review to be analyzed, also improve service of products, and this will help us to improve marketing strategy of any products. Sentiment analysis nay identify the emotions represented in the form of textual data and then emotions are analyzed. Opinion mining may identify the opinions of customers on a particular product. In traditional sentiment analysis, predictions are done at documents level or

sentence level or word level, and calculate frequency of sentiments from documents or reviews. In traditional sentiment analysis assume that one document has one sentiment, and it is practically not possible. The complete process of sentiment analysis of a particular product is shown in figure 1.

Figure 1: Sentiment analysis

Reviews on products are gathered from all its customers. Removing irrelevant sentiments and only relevant sentiments are identified. Choose features of interest, classification of sentiments are done at word or sentence or document level. At the end frequency of each sentiment is calculated. Sentiment analysis is current trending field of research and can be used in many recommendation systems. In this paper, we conducted detailed survey on sentiment analysis with an objective of identifying different classes of machine learning algorithms suitable for sentiment analysis. Identification of proper or correct method of sentiment analysis is critical and also important. The previous methods for machine learning based sentiment analyses are categorized into unsupervised, semi supervised, and supervised learning techniques. Taxonomy of machine learning based sentiment analysis is shown in figure 2.

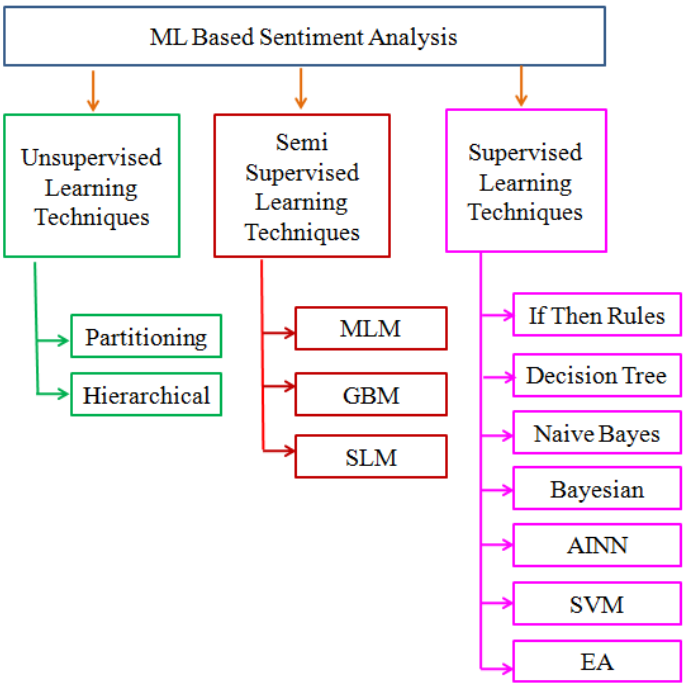

Figure 2: Taxonomy of ML based sentiment analysis

This paper is organized as, section II describes related work on sentiment analysis, section III includes comparative study of unsupervised, semi supervised, and supervised machine learning based sentiment analysis.

Related Work

Authors in [7], described the problem of sentiment analysis with two parameters, one is sentiment, and other one is target. Targets must be fixed by end users, the target is on either any topic or product, sentiments are of customer review on product, and it may be negative or neutral or positive. The data which is used in sentiment analysis are also plays important role, all these data are collected from customers in the form of reviews, all these reviews are analyzed, and based on results business people take decisions [4]. All this reviews are collected from users with help of online websites, and all these websites are main source of data for sentiment analysis. Sentiment analysis is applied to study customer reviews, stocks of different companies, topics, articles, and news debates on a single topic. Now a day's social media is a one of the big source of data where users can share their ideas or sentiments or opinions with others on a topic of interest. Social media data is also used as a dataset in sentiment analysis [13].

In if then rules method, set of rules framed to classify data, LHS of rule is framed using DNF (Disjunctive normal form) with constraints, and RHS is a outcome (class label). More number of conditions is used in rules, most commonly used conditions are support threshold and confidence threshold, support is measure in terms of frequency of an item, confidence is conditional probability between RHS to LHs [6]. In decision tree, tree is constructed with conditions, during training phase system is trained decision tree, and all leaf nodes are treated as class labels. Decision tree classifier for sentiment analysis is a minor change in C4.5 or ID3 algorithm [11].

Authors in [9], used SVM as a classifier to know frequency of sentiments, authors taken opinions as main subject, they applied proposed work on blogs, they extracted user queries in the form of opinions, and all opinions are categorized by using SVM classifier. Authors in [5], used neural network based classifier for sentiment analysis, one input layer, one or

more hidden layer, one output layer, and neurons are distributed in all three types of layers to perform sentiment analysis [10].

Comparative Study

A. Unsupervised Machine Learning Approaches for Sentiment Analysis:

Mostly all sentiment analysis approaches are working based on supervised learning techniques where all data samples must be trained with class labels. In some, cases it is not possible to gather and label data. Samples, mostly in text data samples are a time consuming process. Unsupervised Learning techniques are used to gather data samples study using sentiment analysis. Unsupervised machine learning based sentiment analysis may use clustering techniques, to classify all sentiments into different clusters without mentioning class tables. All unsupervised techniques are broadly classified into either hierarchical or partitioning based clustering.

Partitioning clustering techniques, in this method data samples are arranged into different clusters and each data samples is exactly assigned into only one cluster. To calculate distance between data samples use distance measure (Euclidian), data samples in one cluster are similar and dissimilar to the cluster samples almost all partition techniques are followed k-means to cluster data samples it starts with user defined (K value) number of cluster scan data samples one by one and merge to cluster based on distance, and each time change cluster centroids. In sentiment analysis if data size is large than apply k-means to cluster data samples.

Hierarchical clustering techniques, in this method data samples are partitioned into different clusters by following hierarchy and all this clusters are arranged in a tree. All hierarchal techniques are classified into two categories, one is agglomerative and second one is divisive. In agglomerative, follow bottom to top method, first for small clusters, combine small clusters based on similarity, combine this process, and at the top all data samples are placed in a single cluster. In divisive, follow top down technique, start all data samples as a one cluster, divide cluster into smaller cluster based on similarities, and continue this procedure until cannot be divided further. Advantages and Limitations of unsupervised machine learning approaches for sentiment analysis are tabulated in Table 1.

S.No.	Algorithm	Advantages	Limitations
1	Partitioning	- Simple and scalable - Easy method for	- Handling of clusters and forming of initial

		sentiments analysis. -Applicable for data with big size.	clusters centriod problem. - Low accuracy.
2	Hierarchical	-Cluster good clusters for data samples which are noisy. -Not required to mention number of cluster at initial stages. -Easy and simple to execute / implemented	-Cost of company is high and it is not appropriate if data set size is high. -It is not possible to remove single data sample.

Table 1: Advantages and Limitations of unsupervised ML approaches for sentiment analysis

B. Semi supervised machine learning approaches for sentiment analysis:

This method is used when complex t get class label of data samples, few class labels are fixed at initial stage, which will useful to extract feature classes, and it follows both supervised and unsupervised learning techniques. Authors in [12] introduced a new method of semi supervised to analyze big data on social media, it is combination of SVM and sealing, and their results proved that semi supervised techniques shows better results. Semi supervised techniques broadly classified into three types, Multi Level Model (MLM), Graph Based Model (GBM), and Self Learning Method (SLM). In MLM method, at a time take multiple data samples, and make an agreement between data samples to solve the problem, and data can be classified into clusters based on agreement. All semi supervised technique classifiers may train single time, trained classifier use to labeled classes to classify unlabelled classes of data samples, once classified it is added to trained data, and repeat this process for all data samples with unknown class labels. Authors in [16] use this method to study customer feedback on movie using sentiment analysis.

In GBM method, first data samples are fit into graph, vertices are filled with sentiments, and edges represent relationship between sentiments. Authors in [3] used GBM to study sentiments, sense support count of each sentiment, and based on results they proved that GBM useful to apply in sentiment analysis. In SLM method, it works in two step processes, first step provide training to classifier using initial labeled classes, second step unknown class labeled data samples classified using classifier, once class of data sample identified immediately it is merged to original

trained data set, and repeat this procedure for all data samples. SLM is used in many applications of sentiment analysis to classify data samples. Advantages and Limitations of Semi supervised machine learning approaches for sentiment analysis are tabulated in Table 2.

S.No.	Algorithm	Advantages	Limitations
1.	MLM	- Solve the problem in various ways	- Assumption of independence of all features.
2.	GBM	-Easy to run. - Performance is high when the graph is constructed correctly.	- Performance is based on graph. - if graph is not fir then accuracy and performance is low.
3.	SLM	-Appropriate when few data samples with class labels. -General classifier used in many applications. -Simple.	-Accuracy is low if data at one phase is wrongly classified. -Miss classified data propagate to next step. -Handle data with noise.

Table 2: Advantages and Limitations of Semi supervised ML approaches for sentiment analysis

C. Supervised Machine Learning Approaches for Sentiment Analysis:

SVM & ANN methods are working based on linear regression. All classification of the attributes are done based one line equation $P = q\,X + r$, where P and X are array of values and X is independent and P is dependent, and q, r are constant. More than attribute values are classified based on hyper straight line equation. Support Vector Machine (SVM) used continuous and discrete values and divide data into different groups. It is one of the classification algorithms, used in number of applications, and accuracy of SVM is also high when compared with other algorithms. Authors in [14] described as that SVM is mostly useful classify unstructured data, classification of text and which intern useful to use SVM in sentiment analysis. Authors in [2] used SVM in sentiment analysis to study options of various customers on a movie. As per [1] combination of SVM with other technique called hybrid mode and provided best accurate results.

Artificial Neural Networks (AINN), is a one of the classification method, it is used to extract patterns from data samples, as input data samples, processed using multiple layers, and produced outcomes. AINN, mostly involve neurons in

minimum three layers, one is input layer, secured on is one or more hidden layers, and third one is output layer. Like a network connections all neurons connected with each other to process user inputs at different layers. Authors in[15] conducted experiments on AINN, added more than one hidden layer, used tent data as input, use sentiment analysis to classify data and they identified that their method classify data with less time. Classification is also done based on Bayes theorem, where we are using probability theory, and classify the data sample based on maximum probability value. Naïve Bayes classification is also used in sentiment analysis, where data sample(s) classified into class (t) need to maximize the value of conditional probability $P\ (t\ |s)$.

$$P(t\ |s) = \frac{P\ (t) * P(s\ |t)}{P(s)}$$

Bayesian network is a graph, in which nodes are represented as attributes and edges represent relationship among attributes. In this method each and every node is assumed to be independent and are occurred randomly, the relationships between attributes are calculated using joint probability formula. Bayesian network is used to identify the actual relationships over opinions and comments of customers and same method is used as classification data samples by sentiment analysis. In addition to Bayesian networks use combination of algorithms like random forest algorithm, C4.5, sum, and so on. Entropy analysis (EA), it is not assuming relationships between nodes, conditional probability, calculated over data sample (s) and class label (t), and which must maximize the value of entropy value.

$$E(t\ |s) = \frac{1}{Z(s)} E(\sum_{i} w_i f_{i,t}(s,t))$$

Where Z(s) is a z-score normalization value,

$f_{i,t}()$ is function of data sample to class t,

W_i weight value of function observed value.

$$f_{i,t}(s,t) = \begin{cases} 1 & Number\ of\ data\ samples > 0 \\ 0 & Otherwise \end{cases}$$

Advantages and limitations of supervised machine learning approaches for sentiment analysis are tabulated in Table 3.

S. No.	Algorithm	Advantages	Limitation
1.	If-then rules	-Quickly classify data samples. -Classifier avoids the problem of	-Performance is poor when data is with noise. -Not easy to interprets data

		over fitting.	samples. -Difficult to classify data samples if more number of rules.
2.	Decision Tree	-It is easy to interpret. -Easy to understand. -Handle data samples with noisy.	-Chance of getting problem of over fitting. -Sometimes not stable.
3.	Naive Bayes	-It is easy to run to classifying data Samples. -It requires less time to train dataset.	-Assumption of independent patterns May not be correct in all the cases.
4.	Bayesian Networks	-Required small training data. -Capable to handle data samples with missing values. -It is very simple and easy method even for big size data sample.	-Remaining time is high. -Not appropriate when the data has more number of attributes.
5.	AINN	-Capable to handle complex data samples. -Generalize data samples even though it is Noisy. -Handling high dimensional data. -Running time is fast.	-It is very much difficult to conduct Experiment. -More space is required to run.
6.	SVM	-Kernel mapping in high dimension data reduce working memory space. -It is easy to train data samples. -Accuracy of classification is more. -Handling high dimensional data.	-No probability theorems are used and Inter operability is very low. -Choose right kernel function for data Mapping or otherwise poor results. -It shows low performances when number of classes are increased.

7.	EA	-It is used when prior probabilities values are not known -Extraction rate is more. -Capable to handle data with big size.	-Change of over fitting of data samples.

Table 3: Advantages and Limitations of supervised ML approaches for sentiment analysis

Conclusion

The data which is used in sentiment analysis are also plays important role, all these data are collected from customers in the form of reviews, all these reviews are analyzed, and based on results business people take decisions. Sentiment analysis is a latest trending topic in the area of text mining and sentiment analysis is used to study customer sentiments or opinions on any applications. Sentiment analysis is a one of the statistical method used in NLP and it is used to study opinions, sentiments, reviews, and so on. In this paper, we conducted survey on various unsupervised, semi supervised, supervised based machine learning methods used in the field of sentiment analysis. In addition to that survey also listed various problems faced by each algorithm, advantages of each technique, and limitation each learning methods.

References

Iti Chaturvedi, Erik Cambria, Roy E. Welsch, Francis Herrera, "Distinguishing between facts and opinions for sentiment analysis: Survey and challenges", Information Fusion, Vol. 44, pp. 65–77, 2018.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, Amir Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Information Fusion, Vol. 37, pp. 98–125, 2017.

Fernando Sanchez Rada, Carlos Iglesias, "Social context in sentiment analysis: For- mal definition, overview of current trends and framework for comparison", Information Fusion, Vol. 52, pp. 344–356, 2019,

Erik Cambria, "Affective computing and sentiment analysis", IEEE Intelligent Systems, Vol. 3, No. 2, pp. 102–107, 2016.

Youngseok Choi, Habin Lee, "Data properties and the performance of sentiment classification for electronic commerce applications", Information Systems Frontiers, Vol. 19, pp. 993–1012, 2017.

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, Antonio Feraco, "A Practical Guide To Sentiment Analysis", Vol. 5, pp. 1–196, 2017.

Ana Valdivia, Victoria Luzon, Erik Cambria, Francisco Herrera, "Consensus vote models for detecting and filtering neutrality in

sentiment analysis", Information Fusion, Vol. 44, pp. 126–135, 2018.

Mohammed Erritali, Abderrahim Beni-Hssane, Marouane Birjali, Youness Madani, "An approach of semantic similarity measure between documents based on big data", International Journal of Electrical and Computer Engineering, Vol. 6, No. 5, pp. 2454–2461, 2016.

Marouane Birjali, Abderrahim Beni-Hssane, Mohammed Erritali, "Measuring documents similarity in large corpus using Map Reduce algorithm", 5th IEEE International Conference on Multimedia Computing and Systems, pp. 24–28, 2016.

Francisco Javier Ramírez Tinoco, Giner Alor-Hernández, Jose Luis Sanchez-Cervantes, Beatriz Alejandra Olivares Zepahua, Lisbeth Rodríguez Mazahua, "A Brief Review on the Use of Sentiment Analysis Approaches in Social Networks", pp. 263–273, 2018.

Mika Mantyla, Daniel Graziotin, Miikka Kuutila, "The evolution of sentiment analysis—A review of research topics", venues, and top cited papers, Computer Science Review, vol. 27, pp. 16–32, 2018.

R. Piryani, D. Madhavi, V.K. Singh, "Analytical mapping of opinion mining and sentiment analysis research during 2000–2015", Information Processing & Management, vol. 53, pp. 122–150, 2017.

Fatemeh Hemmatian, Mohammad Karim Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis", Artificial Intelligence Review, Vol. 52, pp. 1495–1545, 2019.

S. Rajalakshmi, S. Asha, N. Pazhaniraja, "A comprehensive survey on sentiment analysis", Fourth International Conference on Signal Processing, Communication and Networking, pp. 1–5, 2017.

Peng Yang, Yunfang Chen, "A survey on sentiment analysis by using machine learning methods", IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference, pp. 117–121, 2017.

Ruijun Liu, Yuqian Shi, Changjiang Ji, Ming Jia, "A survey of sentiment analysis based on transfer learning", IEEE Access, vol. 7, pp.85401–85412, 2019.