IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue-10, May 2019 E-ISSN: 2347-2693

A Survey of Sentiment Analysis based on Machine Learning Techniques

Riya Jain [1*], Siddharth Dutt Choubey[2]

[1] Dept. of Computer Science & Engg, Shri Ram Institute of Science and Technology, Jabalpur, India
[2]Dept. of Information Technology, Shri Ram Institute of Technology, Jabalpur, India

Corresponding Author: riya.jain1313@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7si10.2428 | Available online at: www.ijcseonline.org

***Abstract*—**Internet has become a major part for every individual. More and more users are inclined to share their reviews on internet. This lead to a massive extent of data on web which require analysis so as to become useful. Extracting user's perception from a large dataset of reviews is a difficult task. Sentiment analysis deals at analyzing user's perception from this huge amount of reviews. The idea behind sentiment analysis aims at finding the polarity of text data and classify it into positive or negative. Machine Learning techniques proves to be very helpful in performing sentiment analysis task. This paper presents the survey of main techniques used for sentiment analysis and sentiment classification

***Keywords*—**Sentiment Analysis, Sentiment classification, machine learning, user review's

I. INTRODUCTION

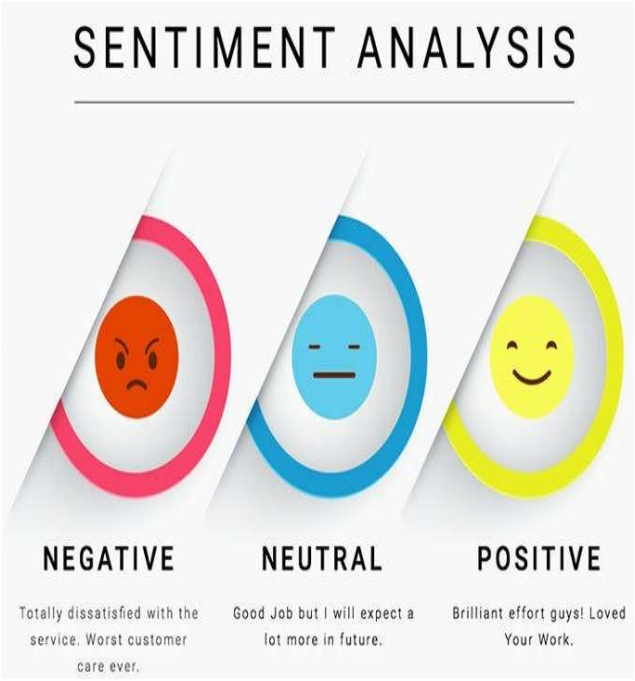
In today's world, with the emerging growth of internet and other web services people tend to share their reviews/opinions online in the form of text. The use of Facebook, Twitter and other social media platforms provide users with the ability to exchange and share their opinions online. Due to the huge amount of data available across various platforms, it is difficulty to analyze user's sentiment and there key points of interest. Sentiment analysis also known as opinion mining is a method of processing and analyzing the subjective text with emotions [1]. The approach employs some Machine Learning techniques which are helpful in constructing a classifier that can classify text based on their emotions. Thus mining of data and classifying reviews based on emotions has become an important field of research.

Sentiment Analysis studies the problem of analyzing text like reviews and post about a person, product, event or any place. Below shows that Sentiment Analysis deals with-

1) Identifying subjective or objective text- Subjecive sentence are those that contains some feelings while objective sentence sticks to the facts
Ex- Subjective – Food was delicious.
Objective – Taj hotel is located at east.

2) Identifying Polarity of the text- based on the subjectivity, sentences can be classified as-

i) Positive (Ex-Delicious food. Service was excellent)
ii) Neutral (Ex-Food was ok but expects a better service quality)
iii) Negative (Ex-Totally dissatisfied with the service)



II. RELATED WORK

Sentiment Analysis is an application of natural language processing and a popular field of research in text mining. Research on sentiment analysis is highly in progress. In this section, a summary of some research works done in the field of sentiment analysis is being discussed-

Upma Kumari et al.[2] proposed the use of Support Vector Machine (SVM) classification algorithm to classify smart phone reviews. They used different datasets for training and testing the model. Calculation of performance evaluation features such as Precision, Recall and F-measure were carried out in order to compute the best accuracy of the product. Experimental results shows that SVM achieves higher accuracy result of 90.99% and it proves to be a robust and most effective one.

In [3], different classification algorithms such as Multinomial Naïve Bayes(MNB), Support Vector Machine(SVM), K nearest neighbor (KNN) and Linguistic regression(LR) are used in order to classify restaurant reviews into positive and negative class. There results show that LR gives the best result with more than 77% accuracy.

Pang et al. employs the use of different machine learning algorithms (Naïve Bayes, Support vector machine and maximum entropy) to divide sentiments into positive and negative categories. They worked on movie reviews. Additionally, they used word bag characteristics frame to carry on the experiment. There work shows that SVM can perform best accurate results in comparison to other techniques. [4].

Abhinash Tripathy et al. [5] used movie reviews of IMDb dataset and employs different supervised machine learning algorithms to classify reviews into positive and negative class labels.

B. Gokulakrishnan et al., analyzed the problem of sentiment analysis and opinion mining of twitter micro blog data. Further they classified tweets as positive, negative and irrelevant classes and analysed the performance of different classifying algorithms [6].

Ghose et al.[7] proposed two ranking mechanisms for ranking product reviews- one is manufacturer- oriented ranking mechanism that ranks according to the effect on sale and second is consumer-oriented ranking mechanism that works according to the consumer expectations. It works as a new concept in recent researches.

III. SENTIMENT ANALYSIS APPROACHES

As shown in the figure, Sentiment Analysis can be classified into machine learning approach and lexicon based approach [8]. In some cases, an approach that combines both machine learning as well as lexicon based approach (hybrid approach) can also be employed.

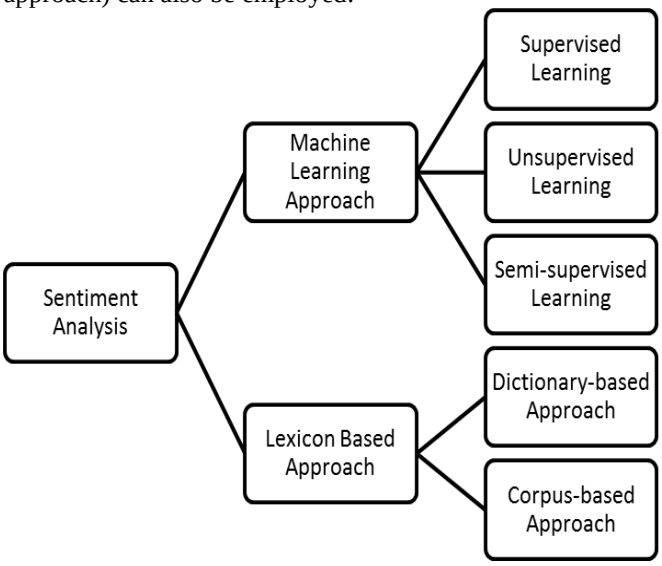


Figure 1 Sentiment Analysis Techniques classification

A. Machine Learning Approach

Machine Learning is the field of study that gives computer the ability to learn without being explicitly programmed. Machine Learning based sentiment analysis aims at employing some machine learning algorithms over the text data for classification. In the next section some machine learning algorithms under supervised and unsupervised learning approaches has been described:

1) Supervised Learning:

In supervised learning, given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output. As shown, supervised learning problems are categorised into "regression" and "classification" problems.

In regression problem, we try to predict results within a continuous output meaning that we are trying to map input variables to some continuous function whereas in classification problem, we try to predict results in a discrete output that is input variables are mapped into discrete categories.

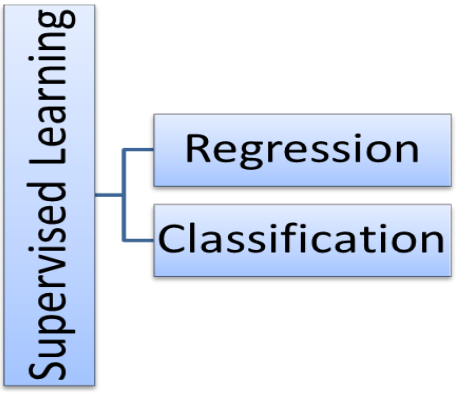


Figure 2 Categories of Supervised Learning

Some most popular supervised learning algorithms are-

i) **Linear Regression**

Linear regression is used to predict output Y from the predictor variable X. In other words we say that in linear regression input values are multiplies by some constants to obtain the output. It creates a correlation between Y and X using a straight line (regression line). The general equation can be written as equation below -

Y = c + aX

where,

Y = Dependent variable (output)

c = constant

a = slope

X = Independent variable

ii) **Naïve Bayesian**

It is a classification technique based on the probabilistic model of Bayes theorem. This model is easy to build and useful for very large data sets. It only assumes the presence of a particular feature in class and is unrelated to any other feature. Posterior probability P(X|Y) can be calculated as equation below-

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Here P(X |Y) is the posterior probability of class (target) given predictor (attribute). P(X) is the prior probability of class. P(Y |X) is the probability of predictor given class. P(Y) is the prior probability of predictor.

iii) **K-Nearest Neighbor (KNN)**

The KNN algorithm is a robust and versatile classifier that is often used as a benchmark for more complex classifiers such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM).In this, our goal is to learn a function h : X → Y so that given an unseen observation x, h(x) can confidently predict the corresponding output y. In the K-nearest neighbor classification algorithm we uses a majority vote between the K most similar instances to a given unseen observation. Similarity is defined according to a distance metric between two data points. A popular choice is the Euclidean distance given by equation below:

$$d(x, x') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \cdots (x_n - x_n')^2}$$

iv) **Support Vector Machine**

It is a classification method which can also work for regression. SVM is suited for text classification as it can handle huge amount of data. Each data item is plot as a point in n-dimensional space (n represents features) with feature value. Feature being the value of a particular coordinate. For example, if data is available with two features like color and height, SVM supports to classify these variables in two dimensional space where each point has two co-ordinates; these co-ordinates are known as Support Vectors.

2) Unsupervised Learning:

Unsupervised learning approach allows us to solve problems with little or no idea what our results should look like. In machine learning, unsupervised learning is a class of problems in which one seeks to determine how the data are organized. As shown, Unsupervised learning problems are categorised into "clustering" and "association" problems.

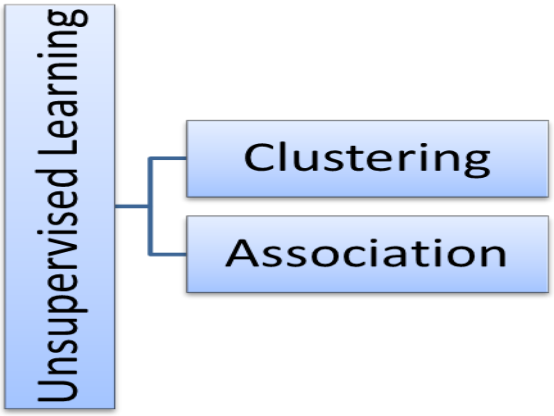


Figure 3 Categories of Unsupervised Learning

i) Clustering

Clustering can be considered the most important unsupervised learning problem that deals with finding a structure in a collection of unlabeled data[9]. Formally, it is the process of organizing objects into groups whose members are similar in some way [8]. Popular techniques under clustering are-

K-means Clustering: The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because different location causes different result. The algorithm steps of K-means are as follows

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid
3. Recalculate the positions of K centroids
4. Repeat Steps 2 and 3 until no changes occur.

- **DBSCAN:** DBSCAN is a density-based clustering algorithm. Primarily, it focuses on finding dense areas and to expand them in order to find clusters[10]. The algorithm is based on two input parameters, Epsilon (Eps), Minimum point (MinPts). [11] Density-based clustering algorithms relates to some important key terms including:
- Core point –It is at the interior of a cluster (the dense area)
- Border point – The neighborhood of a core point having less MinPts within the specified radius
- Noise point – Any point not forming the cluster and even not "absorbed" through expansion is considered as noise point.

3) Semi-supervised Learning:

Semi-supervised is the one that is performed on the dataset in which some data are labeled while most of it are unlabeled.

IV. GENERALISED APPROACH FOR SENTIMENT ANALYSIS PROCESS

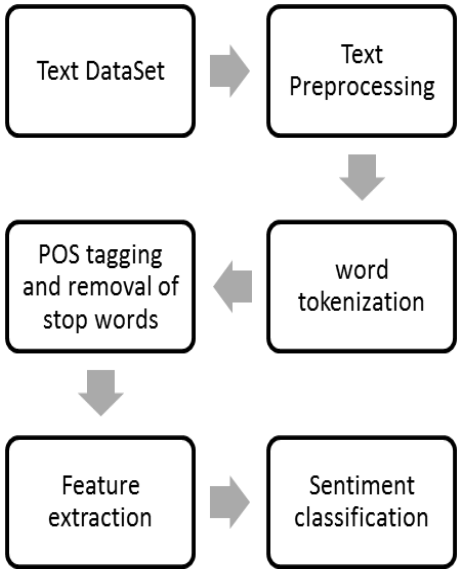


Figure 4 Sentiment Analysis process

Above figure shows the basic process followed for implementing sentiment analysis.

Text processing module includes tokenizing words, parts-of-speech tagging and removal of stop words. Tokenizing is the process of breaking the whole document into separate words so as to make it easy to learn about the sentiments of a user review. POS tagging is the process of reading the text in some language and assign parts of speech to each word such as noun, verb, adjective etc. For eg-"example = 'Movie gives us a rare glimpse into a culture most of us don\'t know ." after applying POS tagging

[('Movie', 'NNP'), ('gives', 'VBZ'), ('us', 'PRP'), ('a', 'DT'), ('rare', 'JJ'), ('glimpse', 'NN'), ('into', 'IN'), ('a', 'DT'), ('culture', 'NN'), ('most', 'JJS'), ('of', 'IN'), ('us', 'PRP'), ('do', 'VBP'), ("n't", 'RB'), ('know', 'VB'), ('.', '.')].

Stop words are those common words which would appear in a text having no meaning or to be of little value. Eg- the, a, to, be etc. Then after pre-processing of the text dataset built model can be applied in order to perform sentiment classification.

V. CONCLUSION AND FUTURE SCOPE

This survey shows a lot of work is being done in the field of sentiment analysis in past few years. By going through many research, it has been observed that Machine Learning techniques had proved to be a great approach in achieving Sentiment Analysis task. Also, we realize that SVM was most versatile and widely used technique of supervised machine learning in the field of sentiment analysis and also it gives higher accuracy results. However, much of the work is performed at classifying sentiments into positive and negative class levels, but for more effective analysis of sentiment or reviews the research can be extended at finding reviews containing some mixed feelings and classifying them accordingly. Thus, the research on sentiment analysis still have a long way to go before reaching the confidence level demanded by practical applications.

REFERENCES

[1] Pang B, Lee L. , "Opinion mining and sentiment analysis" FoundTrends Inform Retriev:1- 135, 2008
[2] Upma Kumari, et al.," Sentiment analysis of smart phone product review using SVM classification technique", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)
[3] F.M Takbir Hossain, Md. Ismail Hossain and Ms. Samia Nawshin."Machine Learning Based Class Level Prediction of Restaurant Reviews" 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)
[4] Pang, Bo, Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques", Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002
[5] Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath, ClassiÞcation of Sentiment Reviews using N-gram Machine Learning Approach, Expert Systems With Applications (2016)
[6] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and A. Perera, "Opinion mining and sentiment analysis on a twitter data stream," in Advances in ICT for Emerging Regions (ICTer), 2012 International Conference on, 2012, pp. 182–188.

[7] Ghose, Anindya, and Panagiotis G. Ipeirotis. "Designing novel review ranking systems: predicting the usefulness and impact of reviews." Proceedings of the ninth international conference on Electronic commerce. ACM,
[8] S. M. vohr et al., "A Comparative study of Sentiment Analysis Techniques", issn: 0975 – 6760| nov 12 to oct 13 | volume – 02, issue – 02.
[9] Introduction to Machine Learning, Second Edition, Ethem Alpaydn, The MIT Press Cambridge, Massachusetts London, England
[10] Christian Bodenstein et. al." Automatic Object Detection using DBSCAN for Counting Intoxicated Flies in the FLORIDA Assay" 2016 15th IEEE International Conference on Machine Learning and Applications
[11] Chetan Dharni and Meenakshi Bnasal. "An improvement of DBSCAN Algorithm to analyze cluster for large datasets" 2013 IEEE International Conference in MOOC, Innovation and Technology in Education (MITE)

AUTHORS PROFILE

Mr. C T Lin pursed Bachelor of Science from University of Taiwan, Taiwan in 2006 and Master of Science from Osmania University in year 2009. He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computational Sciences, Department of Electronic and Communication, University of Taiwan, Taiwan since 2012. He is a member of IEEE & IEEE computer society since 2013, a life member of the ISROSET since 2013, ACM since 2011. He has published more than 20 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. His main research work focuses on Cryptography Algorithms, Network Security, Cloud Security and Privacy, Big Data Analytics, Data Mining, IoT and Computational Intelligence based education. He has 5 years of teaching experience and 4 years of Research Experience.

Mr C H Lin pursed Bachelor of Science and Master of Science from University of New York, USA in year 2009. He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Telecommunication, University of New York, USA since 2012. He is a member of IEEE & IEEE computer society since 2013, a life member of the ISROSET since 2013 and ACM since 2011. He has published more than 20 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. His main research work focuses on Cryptography Algorithms, Network Security, Cloud Security and Privacy, Big Data Analytics, Data Mining, IoT and Computational Intelligence based education. He has 5 years of teaching experience and 4 years of Research Experience.

