Sentiment Analysis Process and Supervised Learning Methods-an Overview

Siji George C G[1], Research Scholar, Dr.B.Sumathi[2], Associate Professor,
Department of Computer Science, CMS College of Science and Commerce

***Abstract*— The subjective information from any written document can be extracted using sentiment analysis. This paper gives an overview of the sentiment analysis process and machine-learning methods. Collecting data, pre-processing the input data, feature extraction and classification are the different stages included in the sentiment analysis process. Supervised machine learning methods such as Naïve Bayes, Support Vector Machine, Maximum Entropy and Neural Network are compared in this paper. Performance can improve by using neural network for sentiment analysis, since it handles tasks that are more complex.**

Keywords— ***Sentiment Analysis (SA), Navies bayes, Support Vector Machine (SVM), Maximum Entropy (MaxEnt), Neural Network***

I.INTRODUCTION

Most of the decision-making process is based on different opinions. In past we had collected different opinions from friends, colleagues, neighbours etc. Now the digital world changed this scenario. As traditional shopping has replaced by online shopping, it is easy to get variety of opinions regarding anything from internet. These opinions are analysed by sentiment analysis. Another term called opinion mining can also be used instead of sentiment analysis.

Sentiment analysis is mainly deals with "what other people think". Sentiment analysis is the process of extracting subjective information from any written document. Machine learning technique and Natural language processing (NLP) are used in sentiment analysis. Detecting the polarity of a given document and emotion recognition are the focuses in sentiment analysis [1].

Sentiment classification can performed in three different levels such as

- Document level
- Sentence level
- Feature level

The first two levels consider the comment/review as a single object for analysis and extract one opinion from one opinion holder. [2]According to the authors, it is better to perform feature level sentiment analysis in which each feature is analysed rather than document/sentence level. The feature level analysis is also called aspect level.

Machine learning uses artificial intelligence (AI). It is used to train the system to learn automatically and improve from experience without being explicitly programmed. The system learned by examples, direct experience or / and instructions, then the system will look for different patterns in data and make good decisions based on the given examples. Machine-learning methods are classified as supervised learning and unsupervised learning.

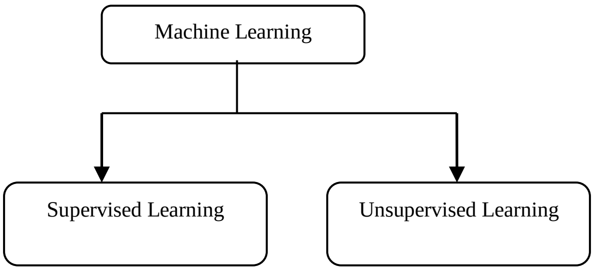

Figure I: Machine-learning methods

Since supervised learning have significant advantages, this paper compared the widely used supervised machine learning methods. This paper is organised as: section II describes the literature review on sentiment analysis and machine learning methods; In Section III, an overview of sentiment analysis process and machine learning methods specially supervised learning methods are described and in section IV comparison among the methods are done. The section V provides the conclusion of the comparison.

II.LITERATURE REVIEW

Bo Pang et al done [3] a comparison among machine learning algorithms used in sentiment analysis. The input movie review data set given to three machine-learning algorithms namely Naïve bayes, MaxEnt and SVM. Within the bag-of-features framework, SVM achieved 82.9% accuracy by using unigram features. In order to produce this result, the author collected 700 positive and 700 negative sentiment documents.

Praniti R. Thanvi et al [4] concentrated on political reviews. The authors used Naïve Bayes algorithm to classify the tweets into Positive, Negative and Neutral by assigning the polarity from -1 to +1. Adverb-Adjective-Verb-Noun (AAVN) combinations helped them to found separate positive, negative results.

Huma Parveen and Prof. Shikha Pandey worked on twitter data set in 2016[5].Their work helped to provide predictions on business intelligence. This work had two sections with considering emoticons and without considering emoticons. Authors observed that the Naïve Bayes algorithm provide more accuracy when pre-processing is done by considering emoticons than without considering emoticons.

Nurulhuda Zainuddin and Ali Selamat used [6] SVM to train a sentiment classifier.The proposed system improved the performance accuracy for the given data set.

A survey on different algorithms used in sentiment analysis is done [7].The authors Vidisha M. Pradhan et al compared both dictionary based approach and machine learning approach. From this survey, it is clear that supervised techniques provide better accuracy compared to dictionary based approach.

Chetashri Bhadanea et al proposed and implemented different techniques for aspect level classification [8]. SVM machine learning algorithm combined with domain specific lexicons algorithms are used to produce the accuracy. 78% accuracy achieved by the proposed system.

Snehal L. Rathod and achin N.Deshmukh introduced a hybrid method [9], which combines sentiment lexicon with machine learning classifier for tweet polarity detection. SVM, and MaxEnt used for 15000 tweets, then 88% of accuracy is achieved.

Anyim Julianne Ankinyi and DR. RobertOboko[10] analyzed sentiment analysis by comparing the performance of the Naïve Bayes Classifier, Maximum Entropy Classifier and Support Vector Machines. The authors observed that feature selection techniques have a great impact on the performance of a classifier. Unigrams, bigrams and trigrams feature selection methods are used. The author concludes that with Trigrams approach, the SVM performed much better than the other two with an accuracy of 82.6%.

G.Vinodhini and RM.Chandrasekaran done [11] a comparison among Back Propagation Neural Network (BPN), Probabilistic Neural Network (PNN) & Homogeneous Ensemble of PNN (HEN). For comparison, they collected data set of product reviews from Amazon review website. The Probabilistic Neural Network (PNN) performs well in classifying the sentiment of the product reviews.

C´ıcero Nogueira dos Santos and Ma´ıra Gatti proposed [12] a system, which uses deep convolutional Neural Network to perform short text sentiment analysis. They worked on both movie review and tweet messages. The proposed network named from Character to Sentence Convolutional Neural Network (CharSCNN) and it achieved a sentiment prediction accuracy of 86.4%.

Apoorv Agarwal et al [13] performed sentiment analysis on twitter data using Part-of-Speech feature selection method. Authors designed a tree representation of tweets.The system use only one representation to combine many categories of features. 60.83% of accuracy is achieved when senti-features is combined with kernel.

III.METHODOLOGY

Polarity identification of a given text is the main task in sentiment analysis. It helps to determine whether the sentiment in a given document is positive, negative or neutral.
Sentiment analysis process can be illustrated in figure (II).

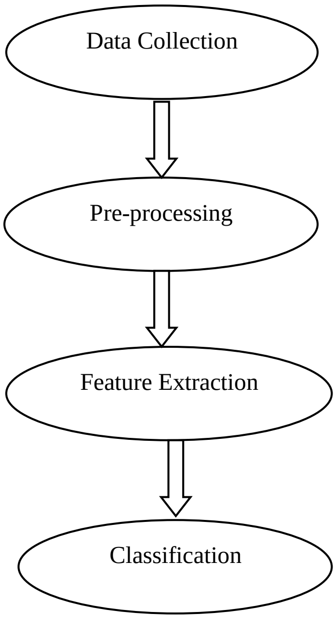

Figure II. Sentiment Analysis Process

Data collection: Data collection is the vital role in sentiment analysis. The result will be based on GIGO (Garbage in garbage out).Careful should be taken for collecting data, otherwise, will reach a wrong result. Widely used data sources for sentiment analysis are Facebook, Twitter, LinkedIn, blogs etc. The collected data will be unstructured. The natural language processing is used to collect and classify the data.

Pre-processing: This is the step, which makes the unstructured data ready for processing. There are lot of irrelevant elements in input and has to be removed. The pre-processing techniques used are

- Tokenization
- Noise removal
- Stemming
- Stop ward removal

In tokenization technique, the sentences or collected reviews are divided into number of tokens. The data collected from internet will have noises such as HTML tag, scripts, advertisement etc. These noises are removed by using pre-processing stage in order to improve the accuracy and performance. Stemming is 'the process of reducing words to their word stem, root or base form'. Consider the example words observe, observes, observer, and observation and all these are stemmed to the root word "Observe". The words that have no value (positive or negative) in a sentiment analysis system are removed in the stop ward removal stage. It include "in"," it""the", "as", "of", "and", "or", "to", etc. Through stop ward removal, the system can save storage space and achieve high performance.

Feature extraction: Feature Selection and extraction is one of the important concepts in machine learning and it affects the performance of the model. New features can be generated from functions of the original features by using the concept of feature extraction.

Classification: After feature extraction, necessary classification algorithms can be applied. Since algorithms in classification determine the accuracy of the polarity detection, classification method selection is very important. The use of particular algorithm is based on available input or data set. Sentiment analysis classifies the orientation of a text in to either positive/ negative. Since algorithms in classification determine the accuracy of the polarity detection, classification method selection is very important. The use of particular algorithm is based on available input or data set.
Researches show that, most of the work done on sentiment analysis uses machine learning approaches. This paper compared four widely used machine-learning methods.

Supervised Machine Learning

In supervised learning, the system will learn from examples. In supervised learning, it is provided with two sets of data-training set and testing set. The system has the capacity to learn from the given training data set so that it can identify the unlabelled examples in the test set with maximum accuracy. Thus the supervised learning process has two steps

- Learning
- Testing

In this paper, the most widely used supervised learning methods are analyzed. They are

- Naïve Bayes
- Maximum entropy
- Support vector machine
- Neural Networks

Naïve Bayes**:** Naïve Bayes is a simple machine learning algorithm.It mainly used for text classification. High dimensional training data set involved in this algorithm. It works based on Bayes theorem. Since it assumes that, the occurrence of a certain feature is independent of the occurrence of other features, it is called Naïve. Membership probability for each class can be predicted by using Baye's theorm. The most likely class will be the highest probability class. The Baye's formula is given below

$$\mathbf{P(c|x)=\frac{P(x|c)P(c)}{P(x)}}$$

Where
P(c|x)- posterior probability
P(x|c)- likelihood
P(c) – class prior probability
P(x) –predictor prior probability

Support Vector Machine**:** Classification problems can be solved by using one of the supervised learning algorithm called SVM. The concept of SVM is very simple. It will create a line, which separates the different classes in a classification problem. This line is used to maximizing the margin between the points on either side of the so-called decision line and this separate line can easily predict the target classes for new cases. The vectors or cases that define the hyperplane are called support vectors.

Consider the training sample $\{(x_i, d_{i,})\}$

where
x_i is the input pattern for the i^{th} example
d_i is the expected output
then the equation for the hyperplane that does the separation is

$$w^T x + b = 0$$

where
x is an input vector
w is an adjustable weight vector
b is a bias
The main aim of the SVM is to find the hyperplane for which margin of separation is maximized.

Maximum Entropy: One of the probabilistic supervised classifier is MaxEnt classifier. It is also called MaxEnt. The class of exponential models are used. Maximum Entropy Principle is related to maximum likelihood. As compared to other algorithms, it makes no independence assumptions. MaxEnt can be used for solving different classification problem. It have only minimum assumptions.

Neural Network**:** Neural networks are used to develop a computer system to perform various computational tasks. They are faster than other traditional system. There are three layers in neural network-input layer, hidden layer and output layer. Each layer consists of one or more nodes represented by circle and the lines between nodes represent the flow of information from one node to other.

The neural network architecture in the context of sentiment analysis can be depicted as

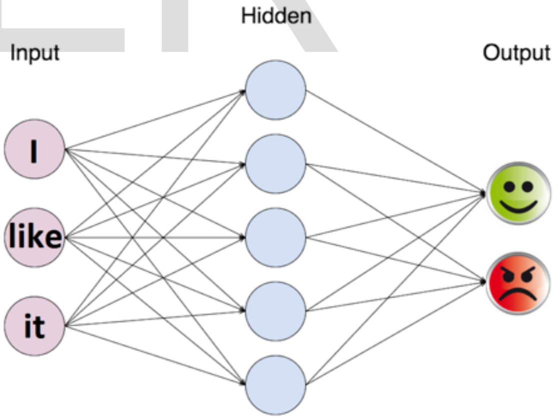

Figure III. Neural network architecture

IV.RESULT

A detailed comparison among Naïve Bayes, SVM, MaxEnt is done in[14].The authors worked on three product review data set and Naïve Bayes algorithm archived high accuracy and it is 81.33%.A useful review of Supervised learning methods is given in [15].According to [16],the selection of classification algorithms should be domain specific. Summary of these analysis is given in the below table.

Table: Comparison of Supervised machine learning methods

Method	Advantage	Disadvantage	Analyzed accuracy
Naïve Bayes	Simple, Fast, High Accuracy, Bag-of-words model, uses small set of training data, scalable, based on assumptions.	Performance decreases as data grows ,can not deals with high dimensional data	81.33%
SVM	Good performance on text categorization, handle real-valued features	Complex, training speed is less and its performance is depends on its parameters	82.7%
MaxEnt	Popular in text classification, Based on empirical data, no biases are introduced, no independence assumptions,performs well with dependent features	Low Performance with independent features	80.2%
Neural Network	Non-linear model, will handle more complicated problems and large amount of data	Generally slower to train, difficult to interpret, performance is sensitive to the size of the hidden layer	86.4%

V.CONCLUSION

Sentiment analysis can perform by using machine-learning methods. Supervised learning methods provide high performance accuracy. Since number of hidden nodes can added in neural network and it handle more complex task, the neural network machine learning method can be used in sentiment analysis. Accuracy can be improved by neural network.

REFERENCES

1. Erik Cambria, Bjorn Schaller Yunqing Xia, Catherine Havasi.”New avenues in opinion mining and sentiment analysis”, IEEE Intelligent System 28(2),15-21,2013.

2. Ms.A.M.Abirami, Ms.V.Gayathri “A survey on sentiment analysis methods and approach”, IEEE Eighth International Conference on Advanced Computing (ICoAC), P 72-76, 2016

3.Bo Pang and Lillian Lee, Shivakumar Vaithyanathan-“Thumbs up? Sentiment Classification using Machine Learning Techniques”- Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, pp. 79-86, July 2002

4.Praniti R. Thanvi, Nikhita S. Sontakke, Shashwati R. Waghmare, Zankhana S. Patel, Prof. Sachin Gavhane,”Sentiment Analysis for Political Reviews using AAVN Combinations”, International Research Journal of Engineering and Technology (IRJET),Volume 5,Issue 01,2017

5.Huma Parveen,Prof. Shikha Pandey-“Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm”, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 5 Issue VI, June 2017.

6. Nurulhuda Zainudhuin,All Selamatt-“Sentiment Analysis Using Support Vector Machine”, IEEE 2014 International Conference on Computer, Communication, and Control Technology,2014

7. Vidisha M. Pradhan,Jay Vala,Prem Balani-“A Survey on Sentiment Analysis Algorithms for Opinion Mining”, International Journal of Computer Applications (0975 – 8887) Volume 133 – No.9, January 2016

8. Chetashri Bhadanea, ”Sentiment analysis: Measuring Opinions”, Elsevier B.V, 45,808 – 814,2015

9. Snehal L. Rathod, Sachin N.Deshmukh,”Sentiment Analysis Using SVM and Maximum Entropy “,International Research Journal of Engineering and Technology (IRJET), **volume 3- issue-8 -august 2016**

10. Anyim Julianne Akiny ,DR. Robert Oboko -“A Comparitive Evaluation Of Sentiment Analysis Techniques On Facebook Data Using Three Machine Learning Algorithms: Naive Bayes, Maximum Entropy And Support Vector Machines”, *School of Computing and Informatics,2014*

11. G. Vinodhini, R.M. Chandrasekaran-“A Comparative Performance Evaluation of Neural network based approach for Sentiment Classification of Online Reviews”, Elsevier B.V,28,2-12,2016.

12. C´ıcero Nogueira dos Santos “Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts”, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 69–78, August 2014.

13.Apoorv Agarval, Boyi Xie-“Sentiment Analysis of twitter”, Proceeding LSM ’11 Proceedings of the Workshop on Languages in Social Media,30-38,January 2011

14. Monali Bordoloi and S.K. Biswas- “Sentiment Analysis of Product using Machine Learning Technique: A Comparison among NB, SVM and MaxEnt”, Volume 118, 71-83,2018

15.Amanpreeth Singh, Narina Takur, Aakanksha Sharma, ”A Review of Supervised Machine Learning Algorithms”,IEEE,1310-1315,2016

16. Hemanth Kumar,”Comprehensive Review On Supervised Machine Learning Algorithms, 2017 International Conference on Machine Learning and Data Science, PP.37-43,2017