Sentiment Analysis based on Different Machine Learning Algorithms

S. Manna

Department of Computer Science and Engineering, Brainware University, Barasat, Kolkata, India

e-mail: *sayanimanna1309@gmail.com*

Abstract— Sentiment analysis is a research topic in the field of text mining. In today's world it plays an important role as we are living in the age of digital world where each and every work is based on internet. These websites are working totally based on online review of various users. Sentiment analysis has gained focus in recent world due to increase in opinion rich web sources such as twitter, online review of products. This paper presents a review of different machine learning algorithms used for Sentiment analysis. A comparative study is being made on decision tree, Naive Bayes Algorithm and Neural Network. Our system is being tested on four products with positive, negative and neutral review. The system processes the text collected as dataset for review and accordingly it is being trained to classify these reviews efficiently.

Keywords— Sentiment analysis, text mining, machine learning, NLP, Decision Tree, Naive Bayes, Neural Network algorithms.

I.   INTRODUCTION

With the advent of Web 2.0 there is an increase in the amount of sentimental content available in the Web [4].These content are found in different social media web sites in the form of movie or product reviews, user comments, testimonials, messages in discussion forums etc. Text Mining is the process of extracting information from a given text. Sentiment analysis is a recent trend of research in the field of text mining. It makes use of Natural Language Processing (NLP), analyses the text, identifies, extracts different emotions (positive, negative or neutral) involved in it. Sentiment analysis is mainly used for analysing emotions from different text document obtained from reviews. Presently sentiment analysis finds huge application in day to day life .For obtaining a correct emotion from a text requires machine learning algorithm which classifies these emotions. This paper makes a comparative study of Decision Tree, Naive Bayes and Neural network algorithms.

I.I Text mining

Text Mining is also known as Text Analytics. It is the process of extracting useful information from a given set of texts. It is useful in the business to find out valuable knowledge from text based content. Text Mining system involves the following technique:

I.I.I Information Extraction

It is used to analyse the unstructured text by finding out the important words and finding the relationships between them. Pattern matching is used to find out the order in text. It transforms the unstructured text into structured form. The Information extraction technique involves language processing modules.

I.I.II Categorization

Categorization technique classifies the text document into one or more category. It includes pre-processing, indexing, dimensional reduction and classification. The text can be categorized using techniques like Naive Bayesian classifier, Decision tree, Nearest Neighbour classifier and Support Vendor Machines.

I.I.III Clustering

Clustering is used to group text documents which are of same type. It has partitions called clusters and each partition will have a number of documents with similar contents. K-means algorithm is often used as clustering technique.

I.I.IV Visualization

Visualization is used to simplify the process of finding relevant information. It helps to display textual information in an unique way.

I.I.V Summarization

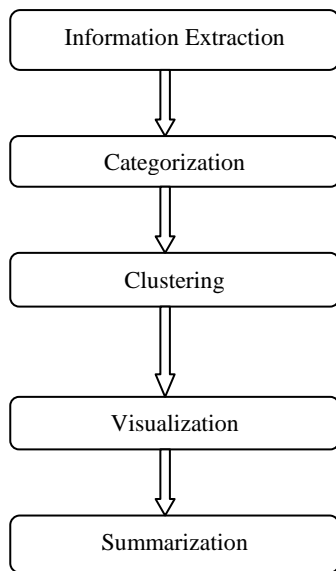Summarization helps to reduce the length of the document and summarize the details of it.

Information Extraction

↓

Categorization

↓

Clustering

↓

Visualization

↓

Summarization

Figure 1. Block diagram of technique of Text Mining

I.II Sentiment analysis

Sentiment analysis determines the attitude of a writer with respect to some text or the overall contextual polarity or emotional reaction to a document, interaction, or event.

I.II.I Classification of Sentiment analysis

It can be classified as document-level, sentence-level, and aspect-level sentiment analysis.

I.II.I.I **Document-level sentiment analysis**

Document-level sentiment analysis considers the whole document a basic information unit and classifies it as expressing a positive or negative opinion or sentiment.

I.II.I.II **Sentence-level sentiment analysis**

Sentence-level sentiment analysis considers each sentence and classifies the sentiment in it. First the sentence is identified as subjective or objective. If the sentence is subjective, Sentence-level SA will determine whether the sentence expresses positive or negative opinions.

I.II.I.III **Aspect-level Sentiment analysis**

Aspect-level SA classifies the sentiment with respect to the specific aspects of entities. The first step is to identify the entities and their aspects.
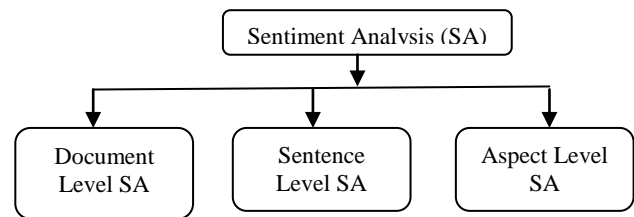
Sentiment Analysis (SA)

↓

| Document Level SA | Sentence Level SA | Aspect Level SA |

Figure 2. Types of Sentiment analysis

I.II.II Different approach of sentiment analysis

I.II.II.I Lexicon based approach

In lexicon based approach, the sentiment of a sentence is calculated by finding lexicon opinion. The text is assigned positive, negative or neutral score based on number of positive or negative or equal number of positive and negative words in a given text [3].The words are categorized as positive or negative based on opinion lexicon. This opinion lexicon can be built in the following approaches.

I.II.II.I.I **Dictionary** based approach

In dictionary based approach, a dictionary is created which is used as reference to categorize words as positive and negative. This dictionary is created first by adding words to it manually. Later on it grows in size by adding synonyms and antonyms of these words which are searched in corpora like WordNet or thesaurus. The disadvantage of this approach is that it becomes more erroneous as its size grows.

I.II.II.I.II **Corpus based approach**

It depends on large corpora for syntactic and semantic patterns of opinion words. The words which are generated are context specific and may require a huge labelled dataset. This corpus based approach may be statistical and semantic.

I.II.II.II **Machine learning based approach**

This approach aims at training machines such that they can recognize the words and accordingly classify them to extract opinion from given text. This learning can be of two types as follows:

I.II.II.II.I **Supervised learning**

In supervised learning, a set of labelled data is provided to train the machine and identify the pattern accordingly.

I.II.II.II.I.I. **Decision Tree**

Decision tree is an algorithm which can be used for classification to train machines and it is in form of tree like structure. ID3 is one of the decision tree algorithms.

1117

I.II.II.II.I.II Linear Classifier:
Linear classifier can be of the following types.

I.II.II.II.I.II.I. Support Vector Machine:
Support Vector Machine (SVM) can be used for both classification and regression challenges. Generally, SVM is used in classification problems. Each data item is plotted as a point in n-dimensional space and the value of each feature is the value of a particular coordinate. After this, classification is done by finding the hyper plane that differentiates the two classes very well.

I.II.II.II.I.II.II. Neural Network:
Neural network is an algorithm which is influenced by human neural network. It tries to imitate the functioning of human.

I.II.II.II.I.II.III. Probabilistic Classifier

I.II.II.II.I.III Naive Bayes:
Naive Bayes Algorithm is used for classification which is based on Bayes algorithm. It starts with an assumption of independence among predictors.

Step 1: Convert the data set into a frequency table.
Step 2: Create Likelihood table by finding the probabilities.
Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

I.II.II.II.II. Unsupervised learning:
In unsupervised learning, no labelled data is provided for training the machine. It learns by experience.

I.II.II.III. Hybrid approach:
It uses combination of lexicon approach and machine learning approach in order to increase efficiency and accuracy of the algorithm.
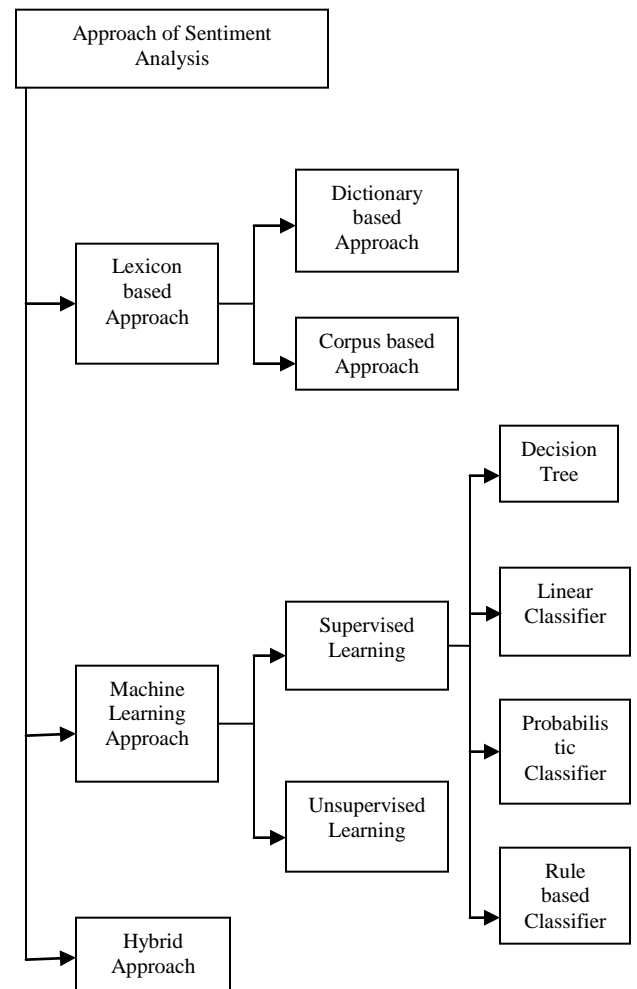

Figure 3 Classification of Sentiment Analysis approach

II.   RELATED WORK

Sentiment analysis is the domain which has developed recently. Analyzing sentiment from a given text uses different approach. Reference [1] describes the role of text pre-processing in sentiment analysis, it also describes that with appropriate feature selection and representation, accuracy in sentiment analysis can be improved using SVM. The classification of sentiment is done using decision tree based feature selection[2].Machine learning algorithm can also be used for sentiment analysis[3].Sentiment analysis can be used in identifying sentiments from different review in different online websites for rating products[6,7].

III.   METHODOLOGY

In this paper we present the sentimental analysis on different comments (review) of users for different products available online. Before the sentiment analysis is performed on these reviews, the data should be converted to proper form and sentiment relevant features are extracted. It involves the following steps:

III.I Data collection

The data for sentiment analysis is collected from online shopping websites which comes as review. These texts (comments of user as review of different product) are collected to form a library which is in form of NLP (Natural Language Processing).

III.II Pre-processing

Step 1: Convert the text to same case (all small letters in our problem).
Step 2: Remove all punctuation mark and special symbols if used (e.g. - comma, colon, semicolon, etc.).
Step 3: Perform Stemming and Lemmatization
Stemming is the process that chops off the end of words and removes the derivational affixes. Lemmatization also does the same thing but with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma*.

III.III Feature Extraction

This is the step where important or required words are identified from the pre-processed text for sentiment analysis. For feature extraction we have considered three models: unigram, bigram and trigram models. From these models a required feature is extracted and a matrix is formed for it to denote the positive, negative and neutral emotions or sentiment of user.

III.III.I Unigram Model

Parts of speech tagging is applied to the text .It identifies monogram or unigram which is extracted to obtain sentiment of an user.

III.III.II Bigram model

This model identifies the feature from text which consists of two words eg. very good or very poor.

III.III.III Trigram model

This model identifies the feature from text which consists of three words eg. not very good .

III.IV Training and testing machine learning classifiers

The extracted features are used to train using a particular machine learning algorithm. In our paper we have used decision tree, Naive Bayes Algorithm and Neural Network for training and a comparison is being made on their performance.

IV.   RESULTS AND DISCUSSION

IV.I Dataset
We have considered some positive, negative and some neutral review on some products available from online shopping sites. For each product, we have obtained reviews belonging to three different categories.

Table 1. Data set of review for different products

Products	No of Positive Review	No of Negative Review	No of Neutral Review
Book	100	100	100
Mobile phone	100	100	100
Television	100	100	100
Laptop	100	100	100

These reviews were then tested after training our model. A matrix is obtained for each category with value +1 for positive, -1 for negative and 0 for neutral review after analysis.

Table 2. Feature matrix of a product indicating its sentiment after analysis.

Product	Actual value	Value after analysis		
Book	Positive	+1	-1	0
	Negative	-1	+1	0
	Neutral	0	+1	-1

Precision: It gives the fraction of retrieved instances that are relevant. It is calculated as

Precision= (No of correct prediction) / (Total no of prediction)

For example: If we are considering a product and its given review is positive. Then number of times its value predicted as correct divided by total number of positive, negative and neutral review.

Precision= (No of times positive value predicted) / (No of times true value predicted + No of times negative value predicted + No of times neutral value predicted)

IV.II Experimental result

We have trained and tested our system with positive, negative and neutral reviews using decision tree, Naive Bayes Algorithm and Neural Network.

Table 3. Comparison of Decision tree, Naive Bayes Algorithm and Neural Network for unigram, bigram and trigram model in terms of precision.

	Decision tree	Naive Bayes Algorithm	Neural Network
Unigram Model	88%	90%	91.3%
Bigram Model	87.4%	89%	89.6%
Trigram Model	76.9%	83.7%	85%

V. CONCLUSION AND FUTURE SCOPE

Sentiment analysis or opinion mining is a topic in the domain of text analysis which analyzes the sentiment or emotion of individual Sentiment of individual are being analyzed from reviews on product available from online shopping sites. The sentiment of these reviews is analyzed by using different machine learning algorithm. .This paper makes a comparison of result using different machine learning algorithm. This is applied on unigram , bigram and trigram model.

In future the sentiment analysis can be performed using deep learning.

REFERENCES

[1] Emma Haddi, Xiaohui Liu, Yong Shi, *"The Role of Text Pre-processing in Sentiment Analysis"*, Information Technology and Quantitative Management (ITQM2013), Procedia Computer Science 17 (2013) 26 – 32

[2] A. Suresh , C.R. Bharathi, *"Sentiment Classifi cation using Decision Tree Based Feature Selection"*, IJCTA, 9(36), 2016, pp. 419-425

[3] Anuja P Jain, Padma Dandannavar, *"Application of Machine Learning Techniques to Sentiment Analysis"*

[4] Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, Sweta Tiwari, " *Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier*"

[5] Kim Schouten and Flavius Frasincar, " Survey on Aspect-Level Sentiment Analysis", IEEE TRANSACTIONS on knowledge and data engineering, vol. 28, no. 3, march 2016.

[6] Alvaro Ortigosa , José M. Martín, Rosa M. Carro, *"Sentiment analysis in Facebook and its application to e-learning",* Computers in Human Behavior 31 (2014) 527–541

[7] Raymond Y.K. Lau , Chunping Li, Stephen S.Y. Liao , *"Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis"*, Decision Support Systems 65 (2014) 80–94.

[8] Zhi-Hong Deng , Kun-Hu Luo, Hong-Liang Yu, *"A study of supervised term weighting scheme for sentiment analysis",* Expert Systems with Applications 41 (2014) 3506–3513

[9] Tahura Shaikh, Dr. Deepa Deshpande, "*Feature Selection Methods in Sentiment Analysis and Sentiment Classification of Amazon Product Reviews"*, International Journal of Computer Trends and Technology (IJCTT) – Volume 36 Number 4 - June 2016

[10] Min-Ling Zhang Zhi-Hua Zhou, *"A Review on Multi-Label Learning Algorithms"*, IEEE TRANSACTIONS ON Knowledge And Data Engineering, VOL. 26, NO. 8, AUGUST 2014

[11] Bo Pang and Lillian Lee, Shivakumar Vaithyanathan, *"Thumbs up? Sentiment Classification using Machine Learning Techniques"*

[12] Arman S. Zharmagambetov, Alexandr A. Pak, *"Sentiment Analysis of a document using deeplearning approach and decision trees"*

[13] Ahmed Alsaffar, Nazlia Omar, *"Study on Feature Selection and Machine Learning Algorithms For Malay Sentiment Classification",* 2014 International Conference on Information Technology and Multimedia (ICIMU), November 18 – 20, 2014, Putrajaya, Malaysia

[14] V.K. Singh, R. Piryani, A. Uddin, P. Waila, Marisha," *Sentiment Analysis of Textual Reviews",* 2013 5th International Conference on Knowledge and Smart Technology (KST)

[15] V. K. Singh, M. Mukherjee & G. K. Mehta, *"Combining Collaborative Filtering and Sentiment Analysis for Improved Movie Recommendations"*, In C. Sombattheera et. al. (Eds.):Multi-disciplinary Trends in Artificial Intelligence, LNAI 7080, Springer-Verlag, Berlin-Heidelberg, pp. 38-50, 2011.

[16] V. K. Singh, M. Mukherjee & G. K. Mehta, *"Combining a Content Filtering Heuristic and Sentiment Analysis for Movie Recommendations"*, In K.R. Venugopal & L.M. Patnaik (Eds.): ICIP 2011, Aug. 2011, CCIS 157, pp. 659-664, Springer, Heidelberg, 2011.

[17] P. Turney, *"Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews",* Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, pp. 417-424, Philadelphia, US, 2002.

[18] Walaa Medhat ,Ahmed Hassan, *"Sentiment analysis algorithms and applications:A survey"* Shams Engineering Journal (2014) 5, 1093–1113

[19] S. Milstein, A. Chowdhury, G. Hochmuth, B. Lorica, and R. Magoulas. *"Twitter and the micromessaging revolution:Communication, connections"* An OReilly Radar Report. 54 pages, November 2008.

[20] J. Ashok Kumar and S. Abirami, *"An Experimental Study Of Feature Extraction Techniques In Opinion Mining*," International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.4, No.1,February 2015.

[21] GurneetKaur and AbhinashSingla, "*Sentimental Analysis of Flipkart reviews using Naïve Bayes and Decision Tree algorithm,"* International Journal Of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. 5, ISSN: 2278- 1323, January 2016.

[22] Aashutosh Bhatt, Ankit Patel, Harsh Chheda and Kiran Gawande, "*Amazon Review Classification and Sentiment Analysis*", International Journal of Computer Science and Information Technologies, Vol. 6, ISSN:0975-9646, 2015.

[23] Lisette García-Moya, Henry Anaya-Sánchez, and Rafael Berlanga-Llavori, *"Retrieving Product Features and Opinions from Customer Reviews",* IEEE Intelligent Systems, vol. 28, pp. 1541-1672, May/Jun 2013.

Authors Profile

Miss. S. Manna pursued B.Tech in CSE from West Bengal University of Technology, Kolkata, West Bengal, India in 2007 and M.Tech in CSE from West Bengal University of Technology, Kolkata, West Bengal, India in the year 2015. Presently, I am currently working as Assistant Professor in Department of Computational Sciences,Brainware University. I am having 3 years of Teaching experience.