Comparative Study of Machine Learning Techniques in Sentimental Analysis

Bhavitha B K
Department of CSE
NMAMIT, Nitte
bhavithabkgowda@gmail.com

Anisha P Rodrigues
Department of CSE
NMAMIT, Nitte
anishapr@nitte.edu.in

Dr. Niranjan N Chiplunkar
Department of CSE
NMAMIT, Nitte
nchiplunkar@nitte.edu.in

***Abstract*—Sentimental Analysis is reference to the task of Natural Language Processing to determine whether a text contains subjective information and what information it expresses i.e., whether the attitude behind the text is positive, negative or neutral. This paper focuses on the several machine learning techniques which are used in analyzing the sentiments and in opinion mining. Sentimental analysis with the blend of machine learning could be useful in predicting the product reviews and consumer attitude towards to newly launched product. This paper presents a detail survey of various machine learning techniques and then compared with their accuracy, advantages and limitations of each technique. On comparing we get 85% of accuracy by using supervised machine learning technique which is higher than that of unsupervised learning techniques.**

Keywords— ***Sentimental analysis, Classifiers, Supervised learning, Unsupervised learning, SVM;***

I. INTRODUCTION

Sentimental Analysis is interpreted as determining the notion of people about distinct existence. Nowadays people are used to review the comments and posts on the product which are known as opinion, emotion, feeling, attitude, thoughts or behavior of the user. Sentimental Analysis is a method for identifying the ways in which sentiment is expressed in texts. Sentimental analysis attempts to divine the posture or notion of a keynoter or author, or author against assertive field or an object. There are many claims in sentiment analysis. First is that, a viewpoint which is treated as positive in one case and will be taken as negative in another case. The next claim is that usually people don't consider their viewpoint in same form. Almost of all reviews incorporate with both positive as well as negative remarks, which can be feasible by interpreting the sentences each at a time. Finding the opinion sites and monitoring them on the web is somewhat difficult. So there will be a need of robotic opinion mining as well as a summarization system.

In sentiment analysis there are three classification levels: document-level classification, sentence-level classification and feature-level sentiment analysis. In document-level classification the main intention is to classify an opinion in the whole document as positive and negative. It speculates entire document as a single unit. The aim of sentence-level analysis is to categorize emotion expressed in respective sentences. In sentence-level the basic step is to recognize the sentence as objective or subjective. Suppose sentence is subjective, it will decide whether it express negative or a positive opinion. In aspect-level analysis it aims to categorize the sentiment in respect of particular entities.

Generally, there are two approaches in sentimental analysis. One is by considering symbolic methods and other one by machine learning method. In symbolic learning technique, which is categorized according to some learning strategies such as learning from analogy, discovery, examples and from root learning. In machine learning technique it uses unsupervised learning, weakly supervised learning and supervised learning. Along with lexicon based and linguistic method, machine learning will be considered as one of the mainly used approach in sentiment classification. The Fig.1 shows the sentiment classification techniques in detail.

1.1 Machine Learning Approach

In artificial intelligence, machine learning is one of its subsections which are proceeding with algorithm that let systems to understand. In machine learning technique it uses unsupervised learning, weakly supervised learning and supervised learning.

1.1.1 Supervised Learning

Supervised machine learning technique associate with the use of a marked feature set to retain some classification function and includes learning of function from the experiment along with its input and output. Supervised learning is task of assuming a function labeled trained data set. Training data set includes set of training examples; each and every example consists of couple of an input data as well as expected output.

1.1.2 Weakly-Supervised and Unsupervised Learning.

In practical these supervised methods cannot be always used, because it needs labeled corpora but they are not available all time. Another option for machine learning is weakly-supervised and unsupervised methods which do not require pre-tagged data. Weakly supervised learning consists of large set of unlabeled data and small set of labeled data. Unsupervised method includes learning device for the input

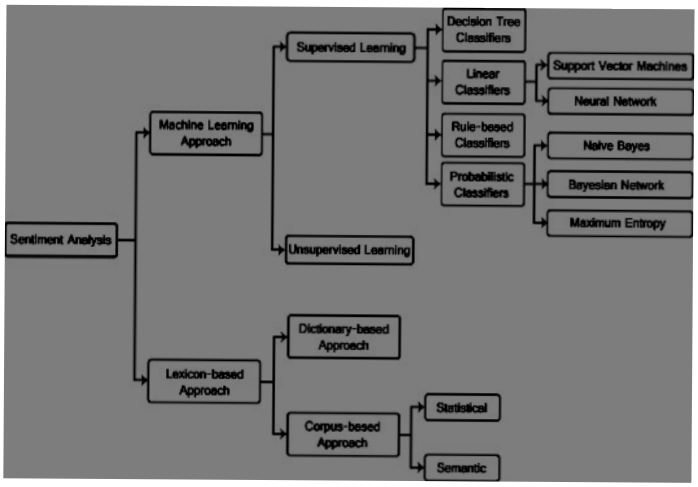

Figure 1 Sentiment classification techniques.

and there is no expected output values are given. Some examples for unsupervised learning approach are cluster analysis, expectation-maximization algorithms. These algorithms use Dictionary based approach to compile sentimental text. A dictionary contains antonyms and synonyms for every word. So this approach is used to find seed sentimental word according to antonym and synonym arrangement of a dictionary. A small set of words is initially collected with known positive or negative coordination. This iterating procedure completed until there is no new words are found.

1.2 Lexicon Based Approach

In lexicon based method it supports a lexicon to achieve sentiment classification through weighting and counting sentiment associated words has to be calculated and labeled. To assemble the viewpoint list there are three major methods are considered: dictionary-based method, corpus-based method and the manual opinion approach.

II. Sentiment Classification Based on Machine Learning Methods

In machine learning technique it uses unsupervised learning, weakly supervised learning and supervised learning.

2.1 Decision Tree Classifier

In Decision Tree classifier, the interior nodes were marked with features and edges that are leaving the node were named as trial on the data set weight. Leaves in the tree are named by categorization. This categories whole document by starting at the root of the tree and moving successfully down through its branches till a leaf node is reached. Learning in decision tree adopts a decision tree classifier as an anticipated model in which it maps information of an item to conclusions of that item's expected value. In decision tree large amount of input can be figure out by using authoritative computing assets in finite time. The main advantages of decision tree classifier are, it is easy to understand and to interpret. This classifier requires small data preparation. But these concepts can create complicated trees that do not generalized easily.

2.2 Linear Classifier

In linear classifier, for classifying input vectors to classes they use linear decision margins. There are many types of linear classifiers. Support vector machine is one of them. This classifier provides a good linear scatters between various classes.

2.2.1 Neural Network

Neural network includes numerous neurons in which this neuron is its elemental unit. Multilayer neural network were used with non-linear margins. The results of the neuron in the previous layer will be given as input for the next layer. In this type of classifier training of data set is more complicated, because the faults must be back-propagated for various layers.

2.2.2 Support Vector Machine

Support Vector Machine (SVM) is known as the best classifier that provides the most accurate results in speech classification problems. They achieved by creating a hyperplane with maximal Euclidean distance for the nearest trained examples. Support Vector Machine hyperplane are completely resolved by a comparatively minute subset of the trained data sets which are treated as support vectors. The remaining training data sets have no access on the qualified classifier. So for the purpose of text classification, the classifier SVMs have been applied successfully and also used in different sequence processing application. SVMs are used in hypertext and text classification since they do not require labeled training data set.

2.3 Rule Based Classifier

As the name indicates in rule based classifiers, data set is designed along with a group of rules. In rules left hand side indicates the condition of aspect set and right hand indicates the class label.

2.4 Probabilistic Classifier

These classifiers use various forms for categorization. This variety of forms takes each and every class as part of that mixture. All various elements are the productive model in which it gives the probability of inspecting a distinct word for that element. These classifiers are also known as generative classifiers. Some of the probabilistic classifiers are Naïve Bayes, Bayesian Network and Maximum Entropy.

2.4.1 Naïve Bayes Classifier

A Naive Bayesian classifier is one of the familiar supervised learning techniques which are frequently used for classification purpose. Their classifier is named as *naive* since it considers the contingency that are actually linked are not depending on the further. Calculation of whole document

feasibility would be the substance in aggregation of all the feasibility report of single word in the file. These Naïve Bayesian classifiers were frequently applied in sentiment categorization since they are having lower computing power when comparing to the other approach but independence assumptions will provide inaccurate results.

2.4.2 Bayesian Network

The main disadvantage of Naïve Bayes classifier is its independent assumption of aspects in data sets. This assumption is the reason for start of using Bayesian Network. This Bayesian network is directed non-cyclic graph where nodes correspond to variables and those edges are correspond to conditional independency. In text classification Bayesian Network is not usually used since it is expensive in computation.

2.4.3 Maximum Entropy

Maximum Entropy classifier is parameterized by a weight set that are used to associate with the joint-future, accomplished by a trained data set by encoding it. This Maximum Entropy classifier appear with the group of classifiers such as log-linear and exponential classifier, as its job is done by deriving some data sets against the input binding them directly and the result will be treated as its exponent.

2.5 K- Nearest Neighbor Classifier

K-Nearest Neighbor is a unsupervised learning algorithm for text classification. In this algorithm the entity is classified with various trained data set along with their nearest distance against each entity. The advantage with this algorithm is its simplicity in text categorization. It also works well with multi-class text classification. The main drawback of KNN is it necessitate with large amount of time for categorizing entities where huge data set are inclined.

In table 1 it shows the comparative observation between different machine learning techniques.

#	Machine Learning Classifier	Advantage	Disadvantage
1	KNN	It is simple and also used for multiclass categorization of document.	It requires more time to categorize when huge number data are inclined. Takes lot of memory for running a process
2	Decision Tree	This is very fast in learning data set. Easy for understanding purpose	It has problem that it is difficult handle data with noisy data Over fitting of data
3	Naïve Bayesian	Simple and work well with textual as well as numerical data. Easy to implement Computationally cheap	Performs very poorly when feature set is highly correlated. It gives relatively low classification performance for large data set. Independent assumption of attribute may lead to inaccurate result.
4	Support Vector Machine	High accuracy even with large data set Works well with many number of dimensions No over fitting	Problems in representing document into numerical vector

Table 1: Comparison between machine learning methods

III. Literature Survey

From the past set of years, many articles, papers and books have been written on sentimental analysis. At the same time some researchers focus more on specific burden like finding the subjectivity expression, subjectivity clues, subjective sentence, topics, and sentiments of words and extracting sources of opinions, while others target is on assigning sentiments to whole document. All analyzers of sentiment analysis have adapted several methods to automatically predict the expression, sentiments of words or a document. The data set for sentimental analysis considered are movie, product review or social media data from the source of internet. They use pattern based approaches, Natural Language Processing (NLP) as well as machine learning techniques.

In paper [6], [7], and [8] authors proposed aspect based sentimental analysis. In the paper [6], during experiment four data sets were used to test the SVM model. Authors have compared Maximum Entropy classifier method for feature extraction with SVM method and they have concluded that SVM Method superior in terms of recall and precision rates. In paper [7], authors proposed different approach which bunch up the benefits of Senti-WordNet, dependency parsing, and co reference resolutions are well organized for the purpose of sentimental analysis. This was done by using Support Vector Machine classifier. The paper [8] presents the comparison between most likely used approaches for sentiments like Naive Bayes, Max Entropy, Boosted trees and Random Forest algorithms. By using Random Forest Classifier in sentimental classification it apparently shows that the result is obtained with greater accuracy and performance. And also this classifier is easy to understand

and the performance would be improved with a time period. Because of aggregation of decision tree the accuracy was improved with higher rate. On behalf of it, the classifier requires high processing power and training time. In the paper they conclude that, if accuracy has the first consideration then Random Forest classifier must be preferred even though it uses high learning time. Due to lesser processing condition and small memory usage the Naïve Bayesian classifier was applied. Alternatively the Max Entropy classifier is used because it requires smaller training time with large memory and processing time. From these papers we can conclude that support vector machine yields higher accuracy in classification of product reviews. But authors have not dealt with sarcastic sentences and comparative sentences.

In the paper [9], [10] and [19], a movie review is analyzed by linking machine learning application with Natural Language Processing technique. In paper [9], authors applied SVM and Naïve Bayes classifier in analyzing the movie sentiments. By this categorization they conclude that linear Support Vector Machine outperforms the Naïve Bayesian in case of accuracy. In the paper [10], authors demonstrate how machine learning technique was used to understand the Malayalam movie comment. For classifying the sentiments two machine learning approaches are used; they are Support Vector Machine and CRF along with rule based approach. In [19] , author compared two most frequently used supervised machine learning approaches SVM and Naive Bayes for sentiment classification of reviews. The result shows that SVM has misclassified more number of data points as compared to Naive Bayes and Naive Bayes approach outperformed the SVM when there are less number of reviews. Authors suggested that there will be a considerable scope of improving in the creation of corpus and effective preprocessing and feature selection Researchers are still working for the automated analysis of score and rating of the movie reviews.

In paper [11],[12],[13] and [14] authors describe the various tools used in sentimental analysis of twitter data. Since the opinions in the twitter are heterogeneous, highly unstructured and along with these it includes positive, negative or neutral in different situation, it is important analyze the sentiments. In the paper [11] authors used lexicon-based methods for classification but it requires small effort in individual labeled text document. In the paper [12] they have shown the outline of recommended methods along with its most recent advancements in the same field. As a result, authors concluded that unsupervised machine learning techniques fails to provide better achievement in sentiment classification than that of supervised learning. In paper [13], they describe the various tools used in sentimental analysis and some approaches for text classification. In this method they use hybrid approach which uses the aggregation of both lexicon based and machine learning techniques. This compound approach then leads to obtain higher classification performance. The fundamental usage of machine learning is its capability to change and to bring qualified design for exact purpose together with its content. In the paper [14], authors proposed Naïve Bayesian classifier to analyze the sentences. Their experimental result shows that Naive Bayesian classifier model which has acceptable achievement for distinct Social Network Site and for large data set in which it consists of long comments.

The challenges for automated analysis of tweets are (i) a single word are considered as subjective in one case and the same word will be treated as objective in another case (ii) same sentence with different discipline (iii) sarcasm sentences (iv) in some case a whole sentence will not be considered because only little part of the text gives the complete contention (v) negative word can be expressed in distinct way in contrast to words like never, no, not etc. Analyzing such contradiction is challenging. That means the twitter analysis still more improved by considering these challenges.

In the paper [15],[1],[5] and [16], authors discussed about existing models for analyzing sentiments of unorganized data which were posted on social media. Analyzing sentiments it doesn't consider objective sentences. Authors proposed approach for sentence classification or sentence of documents. For this purpose [15][1] used SVM, Naive Bayes, Part of Speech and SentiWordNet techniques. From the result they conclude that machine learning classifier for instance Naïve Bayesian and Support Vector Machine yield the highest efficiency. And also act as basic standard model for all classification. But lexicon approach is very aggressive in sentiment. For this problem deep learning approach was introduced. By comparing lexicon based approach with machine learning, the two classifiers such as SVM and Naive Bayes provides higher accurate values in classification.

The paper [5],[16], depicts that for the purpose of sentiment analysis they use classifier of Support Vector Machine (SVM) on the benchmark feature sets to scale the sentiment classifier. To extract the classical features of data set, weighting scheme like N-grams and other weighting scheme were applied. For selecting requested feature to the classification they go into the Chi-Square weight feature. In the present method the structure involves preprocessing, aspect selection, aspect extraction and finally the data sets are classified. Since SVM is having great potential to hold big data set, the text classification is done with good result. Other mentionable advantage is SVM is robust with sparse set of examples. N-gram, unigram and other weighting schemes are input to the SVM classifier. Based on these weighting schemes some standard data sets are routine to train the classifier. In the experimental result it found that unigrams outperforms bigram and n-gram model. To improve the accuracy in classification authors suggested using Chi-Square aspect selection scheme.

In paper [17], the author presents comparative analysis of currently used techniques for sentiment analysis which includes lexicon-based and machine learning techniques

along with cross-lingual and cross-domain method. Finally they conclude that machine learning approach provides highest accuracy and lexicon-based approach are highly competitive and also manually needs more effort in labeling document.

Table 2 shows machine learning approach SVM yields highest accuracy as compared to Naïve Bayes and Senti-WordNet.

	TP	FP	FN	TN	Accuracy
Senti-WordNet	148	91	52	109	64.25%
NB	156	81	44	119	68.75%
SVM	135	51	65	149	71.00%

Table 2: Performance comparison of learning methods with Senti-WordNet

Below table 3 shows performance contrast between different sentiment classification approaches.

	Method	Data set	Accuracy
Machine learning	SVM	Movie reviews	86.40%
	CoTraining SVM	Twitter	82.52%
	Deep learning	Standard benchmark	80.70%
Lexicon based	Corpus	Product Reviews	74.00%
	Dictionary	Amazon	---
Cross-lingual	Ensemble	Amazon	81.00%
	Co-Train	Amazon, IT168	81.30%
	EWGA	IMDb moview review	>90%
	CLMM	MPQA, NTCIR,ISI	83.02%
Cross-domain	Active learning	Book, DVD, Electronics, Kitchen	80% of average
	Thesaurus		
	SFA		

Table 3: Performance comparison of sentiment classification technique

In paper [18], they have taken online movie reviews for analyzing sentiments. For classification they used three supervised learning approaches such as Naïve Bayes, SVM and kNN. Experimental results show that SVM method beat the kNN and Naïve Bayes approaches. Table 4 shows the collected reviews for sentiment classification.

# experiment	Positive	Negative	Total
1	50	50	100
2	100	100	200
3	150	150	300
4	200	200	400
5	400	400	800
6	550	550	1100
7	650	650	1300
8	800	800	1600
9	900	900	1800
10	1000	1000	2000

Table 4: Collected reviews

The accuracy obtained by using three algorithms are shown in table 5. They have done 10 experiments for each approach. Result shows that even data is either small or large SVM provides higher accuracy than NB and kNN.

# experiment	# reviews	SVM (%)	Naïve Bayes (%)	kNN (%)
1	50	60.07	56.03	64.02
2	100	61.53	55.01	53.97
3	150	67.00	56.00	58.00
4	200	70.50	61.27	57.77
5	400	77.50	65.63	62.12
6	550	77.73	67.82	62.36
7	650	79.93	64.86	65.46
8	800	81.71	68.80	65.44
9	900	81.61	71.33	67.44
10	1000	81.45	75.55	68.70

Table 5: Accuracy obtained after testing data set

IV. Conclusion

This paper includes outline of current works that done on sentimental classification and analysis. From the survey we can conclude that supervised learning methods like Naive Bayesian and Support Vector Machine are considered as standard learning method. Support Vector Machine provides excellent accuracy as compared to many other classifiers. In terms of accuracy we concluded that with small feature set Naive Bayes performs well, if large feature set is taken then SVM will be the best choice. Lexical based approaches are ideally aggressive because it requires manual work on document. Maximum Entropy also performs better but it is suffered from over fitting. Many researches implemented opinion mining different techniques but still there is a need of automated analysis which addresses all the challenges of sentimental analysis simultaneously. A more innovative and effective techniques required to be invented which should overcome the current challenges like classification of indirect opinions, comparative sentences and sarcastic sentences.

REFERENCES

[1] Hailong Zhang, Wenyan Gan, Bo Jiang, "Machine Learning and Lexicon based Methods for Sentiment Classification: A Survey",978-1-4799-5727-9/14 $31.00 © 2014 IEEE.

[2] Walaa Medhat a,*, Ahmed Hassan b, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal (2014) 5, 1093–1113.

[3] Xing Fang* and Justin Zhan, "Sentiment analysis using product review data", Fang and Zhan Journal of Big Data (2015) 2:5 DOI 10.1186/s40537-015-0015-2

[4] Kaijie Guo, Liang Shi*, Weilong Ye, Xiang Li, "A Survey of Internet Public Opinion Mining", 978-1-4799-2030-3 /14/$31.00 ©2014 IEEE.

[5] Nurulhuda Zainuddin, Ali Selamat, "Sentiment Analysis Using Support Vector Machine", 978-1-4799-4555-9/14/$31.00©2014 IEEE.

[6] Chuanming Yu, "Mining Product Features from Free-Text Customer Reviews: An SVM-based Approach", iCISE 2009 December 26-28, 2009, Nanjing, China.

[7] Raisa Varghese, Jayasree M, "Aspect Based Sentiment Analysis using Support Vector Machine Classifier", 978-1-4673-6217-7/13/$31.00_c 2013 IEEE

[8] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis", International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6261-6264.

[9] Gautami Tripathi1 and Naganna S, " Feature Selection And Classification Approach For Sentiment Analysis",Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.2, June 2015.

[10] Deepu S. Nair, Jisha P. Jayan, Rajeev R.R, Elizabeth Sherly, "Sentiment Analysis of Malayalam Film Review Using Machine Learning Techniques", 978-1-4799-8792-4/15/$31.00_c-2015-IEEE

[11] Vishal A. Kharde, S.S. Sonawane , " Sentiment Analysis of Twitter Data: A Survey of Techniques",International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016.

[12] Neha S. Joshi, Suhasini A. Itkat, "A Survey on Feature Level Sentiment Analysis", International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5422-5425.

[13] Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni, Tiziana Guzzo ,"Approaches, Tools and Applications for Sentiment Analysis Implementation",International Journal of Computer Applications (0975 – 8887) Volume 125 – No.3, September 2015.

[14] Shun Yoshida, Jun Kitazono, Seiichi Ozawa, Takahiro Sugawara, Tatsuya Haga and Shogo Nakamura, "Sentiment Analysis for Various SNS Media Using Naive Bayes Classifier and Its Application to Flaming Detection",978-1-4799-4540-5/14/$31.00 ©2014 IEEE .

[15] Jalaj S. Modha*, Prof & Head Gayatri S. Pandi, Sandip J. Modha, " Automatic Sentiment Analysis for Unstructured Data", IJARCSSE Volume 3, Issue 12, December 2013.

[16] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in twitter events," J. Am. Soc. Inf. Sci. Technol., vol. 62, no. 2, pp. 406–418, Feb. 2011.

[17] Hailong Zhang, Wenyan Gan, Bo Jiang, "Machine Learning and Lexicon based Methods for Sentiment Classification: A Survey", 978-1-4799-5727-9/14 $31.00 © 2014 IEEE DOI 10.1109/WISA.2014.55.

[18] P.Kalaivani, Dr. K.L.Shunmuganathan, "Sentiment Classification Of Movie Reviews By Supervised Machine Learning Approaches", ISSN : 0976-5166 Vol. 4 No.4 Aug-Sep 2013.

[19] Suchita V Wawre1, Sachin N Deshmukh2 , "Sentiment Classification using Machine Learning Techniques", International Journal of Science and Research (IJSR) Volume 5 Issue 4, April 2016.

[20] Walaa Medhat a,*, Ahmed Hassan b, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal (2014) 5, 1093–1113.

[21] Pratiksha Y. Pawar and S. H. Gawande, "A Comparative Study on Different Types of Approaches to Text Categorization", International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012.

[22] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica 31 (2007) 249-268 249.