ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Sentiment Analysis and Machine Learning Based Sentiment Classification: A Review

Poonam Choudhari
Department of CSE, AISECT University
Bhopal,India

Dr. S. Veena Dhari
Department of CSE, AISECT University
Bhopal,India

Abstract : Sentiments are the feelings expressed by an individual about entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. Understanding the meaning of sentiment and interpreting them in positive, negative and neutral classes in a automated way is known as *Sentiment Analysis*. Machine Learning Classifier algorithms are used to classify the sentiment in different classes. This paper presents a comprehensive review of sentiment analysis addressing different concepts in this area, challenges applications along with a list of research areas in this field. It also addresses major machine learning algorithms used for sentiment classification, their comparisons and recent research work in this area.

Keywords : Opinion, Sentiment Analysis; Naïve Bayes ; Maximum Entropy; Decision Tree; Support Vector Machine.

I. INTRODUCTION

Opinions of people about some product or topic play an important role as a deciding factor in our life which influences our perception to a great extent and thus help in taking decisions. For example before deciding to purchase a product an individual take opinions of others who have used that product or having some knowledge about the product. Companies make opinion polls to take customer feedbacks before launching some product or about the products already in market, capturing these insights about customer behavior and preferences that could help generate more revenues. With the advent of web ,people have started expressing their views/opinions on social platforms and today a large amount of opinionated web data is available. Most of this user generated content are unstructured and opinions are hidden in the long written posts, which makes it difficult for a reader to analyze and get useful information from such textual data. The large amount of data that these social networking sites generate cannot be manually analyzed. Thus there is need for some automated analyzing technique which is known as *Sentiment Analysis*. Sentiment Analysis, also called *Opinion Mining*, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes.[1] That means Sentiment Analysis in simple terms is the field where machines are taught to understand human sentiment.

Sentiment analysis is different from traditional text categorization. In traditional text categorization is done by topic, there can be many classes that are user and application-dependent for a given document i.e. there can be only two classes or as many as hundreds or thousand of classes according to different applications. In Sentiment classification ,there are relatively few classes i.e. binary classification(positive ,negative),degree of positivity, star rating etc. that generalizes across many domains[2].This paper presents a survey related to Sentiment Analysis ,knowing which further research work can be done in this field.

II. SENTIMENT ANALYSIS CHALLENGES

There are many challenges in sentiment analysis that are to be faced during their mining due to the content of text messages in a unstructured manner(that makes it more difficult from traditional text categorization). They are as follows[3][4] :

- The language used in text for expressing feelings/emotions may be informal, not in proper work language, may contain words that carry less meaning and are in ungrammatical manner.
- High volumes of messages are posted everyday with a wide array of topics and vocabulary that are not domain specific. The length of messages are also variable i.e. may contain one word or may contain large array of words .
- The messages may contain internet slang, abbreviations , shortenings of words, non-conventional spellings , special strings, hash tags, emoticons, meta information , hyper links ,likes, locations are some of the complexities that need to be addressed.
- The messages may contain sarcasm .It means the text may contains the sentiment opposite of what user is writing like" What a great tool! It stopped working after two days."
- The problem of Word Sense Disambiguation may be faced in the text i.e. meaning of same word changes with respect to context like for example the sentence "The activity really sucks" is negative, whereas the sentence "The vacuum cleaner sucks really good" uses the same verb but in a positive way.
- The problem of finding Opinion spam is also a big challenge . It refers to "illegal" activities that try to mislead readers by giving undeserving false

positive opinions to some target entities in order to promote the entities and/or by giving false negative opinions to some other entities in order to damage their reputations.

III. RESEARCH AREAS OF SENTIMENT ANALYSIS

The following areas can be taken into consideration for research[1][5] :

- The different levels of sentiment analysis i.e. sentence level, document level and aspect level can be considered for research purpose.

- The research in this field can be focused on different languages based sentiment analysis and domain based sentiment analysis i.e. with respect to different domains like product reviews, hotel reviews, movies reviews ,political reviews, health care related reviews can be considered .

- The research in this field can be focused on to determine subjective/objective polarity i.e. whether the text contains some factual information (objective text) or some opinion(subjective text) is expressed (which is called as Subjectivity Classification).

- The performance of the algorithm used for sentiment analysis in terms of speed and accuracy can be improved. This can be done by focusing on the different steps involved in the algorithm of sentiment analysis like Pre-processing step, feature-selection step etc. Data Pre processing techniques play a important role in improving the performance of algorithms, so research can be done on improving the conventional preprocessing techniques. After Pre-processing, feature selection and feature reduction is also a important step as the features selected play a important role in overall performance of algorithm.

- The approach used for Sentiment Classification i.e. Machine learning(Supervised and Unsupervised), Lexicon based ,or Hybrid of two, can be considered .So work can be done on Sentiment Classification so as to improve the performance of system.

- The visualization of the output can be considered i.e. opinion summarization expressed in terms of binary(i.e. positive or negative),ternary(positive, negative, or neutral),degree (i.e. strongly positive, weakly positive, strongly negative, weakly negative ,star rating etc).

- The difficulties that often occurs when dealing with textual comments can also be considered like opinion sarcasm, word sense disambiguation and opinion spam that were discussed previously.

IV. APPLICATIONS OF SENTIMENT ANALYSIS

Sentiment analysis has grab the attention of the researchers with the rapid increase of possible applications. The major applications of Sentiment Analysis are the following[2]:

- Evaluating customer satisfaction metrics: In present day scenario world is running on social media like Twitter, Face book, and Amazon (Reviews) ,the customers would not hesitate to take social media to express their opinion on whether they are satisfied with the product/service or not , difficulties they are experiencing ,and new features they would like to see. The new customers can compare the different brands and judge whether to buy the product or not, by analyzing ratio of the positive/negative customer reviews . The company can also get to know about their product performance in the market and improve their marketing strategies by analyzing the online customers reviews through sentiment analysis.

- Stock market Prediction: To investigate the information about stocks and identify current stock price trends .Sentiment analysis can be done on the stock market reviews given by experts and consumers .

- Government policy tracker :The government can get to know about people views whether their policies are effective or not ,or the pros and cons of their policies implemented or future policies to be implemented. For e.g. the Obama administration used sentiment analysis to get to know about public opinions to policy announcements and campaign message ahead of 2012 presidential elections.

- Recommendation Systems: When the number of online review of a product is available on large scale, summarization is used. Opinion mining of these large number of product reviews can provide effective summarized information by classifying the people's opinion about the products into positive and negative, the system can say which one should get recommended and which one should not get recommended.

V. MACHINE LEARNING BASED SENTIMENT CLASSIFICATION

The major machine learning sentiment classification algorithms like Naïve Bayes, , Decision tree, Maximum Entropy and SVM are defined in a brief manner below. Also the advantages and limitation of each algorithm with a summarized comparison is also defined ,so that knowing which according to the defined characteristics, the Machine Learning algorithm can be utilized to improve the performance of sentiment classification.

A. *NAÏVE BAYES*

Naive Bayes classifier[6][7] is a simple probabilistic classifier based on Bayes theorm. In Naïve Bayes technique , the basic idea to find the probabilities of class labels given a text document by using the joint probabilities of words and class labels. This classifier is termed as naïve as it believe on assumption of word independence that each feature i.e. words have no dependency or connection with other words in the document/sentence being considered for classification purpose. It means that all attributes are independent such as one word does not affect the other in deciding whether or not the tweet or review is positive, negative or neutral.

The Bayes theorem for finding probability, is defined, for a given data point x(i.e .word) and class 'c'(here in case of SA ,c = positive ,negative or neutral):

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

P(c|x) =Posterior probability(i.e. the resultant probability of attribute x in test set belong to class label 'c')

P(x|c)= Maximum Likelihood or Conditional Probability(i.e. overall probability of attribute x in training set belong to class label 'c')

P(c)=Class Prior Probability(i.e. probability of Class 'c' in training set)

P(x)=Predictor prior probability (i.e. probability of attribute x in all the class labels in training set)

1)Advantages of Naïve Bayes :

- It is simple to implement .i.e. easier to predict class label on test data.
- Training time and prediction time required is less as compared to other text classification algorithms.
- If textual data for training is fairly little, then high bias/low variance classifier i.e. Naïve Bayes classifiers does well in such circumstances. As the interactions between the attributes are ignored in the model, there is no requirement of examples of these interaction and therefore less data is required than other text-classification algorithms.
- The performance of classifier is good with independent feature vectors.

2)Disadvantages of Naïve Bayes :

- The performance of algorithm can degrade if the data contains highly correlated features.
- One of the problem encountered in Naïve Bayes is Zero Observation problem .i.e. if a categorical attribute has a value in the test set that was not there in the training set .Then the model will assign a zero probability and be unable to make a prediction .But Some techniques like laplacian smoothing can be applied to overcome zero observation problem.

B. *DECISION TREE*

Decision Tree[8][9]is a non-parametric supervised method used for sentiment analysis text classification. The motive is to develop a tree-like structure model that predicts the value of target class in this case of sentiment analysis i.e. positive, negative or neutral by learning simple decision rules obtained from trained data features set.

Decision Tree classify instances or examples by starting at the root of the tree and moving through it until a leaf node. The selection of an attribute word to test at each node i.e. splitting node – choosing the most useful attribute for classification which gives maximum information. For example if ID3 algorithm is used for decision tree implementation then the best split is chosen by using information gain .Information gain measure based on entropy, is used to select the best split among the candidate attribute at each iteration i.e. the attribute which measures how well a given attribute separates the training data according to target data.Entropy provides an informative –theoretic approach to measure the goodness of a split.

If a set of records T is partitioned into set of disjoint classes $p_1, p_{2,.....}, p_n$ on the basis of the value of class attribute, then the information need to identify the class of an element of T is:

Info(T) = Entropy(P),where P is the probability distribution $p_1, p_{2,.....} p_n$.

$$\text{Entropy}(P)=-[p_1\log(p_1) + p_2\log(p_2) +\ldots.+ p_n \log(p_n)\ ]$$

The information gain due to a split on attribute X as

$$\text{Information Gain}(X,T) = \text{Info}(T) - \text{Info}(X,T)$$

$$\text{where Info}(X,T)=\sum_{i=1}^{n} \frac{|T_i|}{|T|}\ \text{Info}(T_i)$$

The information gain represents the difference between the information needed to identify an element of T and the information needed to identify T after the value of attribute X is obtained.

1)Advantages of Decision Tree

- Simple to understand ,visualized and to interpret even by non programmers. Decision trees are white-box classification algorithm means they are able to generate understandable rules in human readable form.
- Simplifies complex interaction between input variable and target output by dividing the original input variables into significant sub-groups.

- Decision tree gives a clear indication of which are the attributes feature i.e. words are most important for classification .
- Decision tree are non-parametric means no specific data distribution is necessary.
- It can easily handle feature interactions and the are robust to outliers.

2) Disadvantages of Decision Tree

- Decision tree learners can create over-complex trees that may not generalizes well This is called over fitting problem. The over fitting decision tree require more space and other computational resource. Pruning helps in handling the over-fitting problem.
- It does not work well with continuous attribute as compared to categorical one.

C. *MAXIMUM ENTROPY*

Maximum entropy also known as Max Ent or multinomial logistic regression is introduced first by Berger and Della Pietra at 1996 is a probabilistic classifier[10][11].This classifier is based on the principle of Maximum Entropy. Maximum entropy model all that is known and assume nothing about what is unknown. This algorithm gives an idea about probability distribution not just probability as in the case of Naïve Bayes . In this algorithm the probability that a document belong to particular class must maximize the entropy.

The mathematical formula that the probability of a document 'd' belong to a particular class c_j is given by:

$$P(c_j|d)= \frac{1}{Z(d)} \exp(\sum_i \lambda_i f_i(d,c_j))$$

Where each $f_i(d,c_i)$ is a feature, λ_i is a parameter to be estimated Z(d) is the normalizing constant that is obtained by summing overall $P(c_j|d)$ over all values of j.

$$Z(d)= \sum_n \exp(\sum_i \lambda_i f_i(d,c_j))$$

1) Advantages of Maximum Entropy

- Perform well with dependent features.
- Ability to combine different kind of statistical dependencies in one uniform framework.

2)Disadvantages of Maximum Entropy

- The Performance degrades when the feature vectors applied are independent.
- The Max Entropy requires more time to train comparing to Naive Bayes, primarily due to the optimization problem that needs to be solved in order to estimate the parameters of the model.

D. *SUPPORT VECTOR MACHINE*

The Support Vector Machine is a non-probabilistic linear classifier. In this, the idea is to non-linearly map the data set into high dimensional vector space and use a linear discriminator to classify the data. SVM is based on the Structural Risk Minimization principle from computational learning theory[12][13].In structured risk minimization, instead of minimizing the error ,minimization is done on upper bound on the generalization error . SVM classify the text by drawing a separating line or the hyper-plane on the scatter plot between positive and negative class label .The document representatives which are nearest to the hyper plane are called Support Vector. Here the goal of SVM is to find a optimal separating Linear hyper plane which maximizes the margin between the two class labeled data pointswhich is shown in the figure. Margin is the distance of the closest point of each class from the separating hyper –plane. In order to calculate the margin ,two parallel hyper planes are constructed on each side of hyper plane.

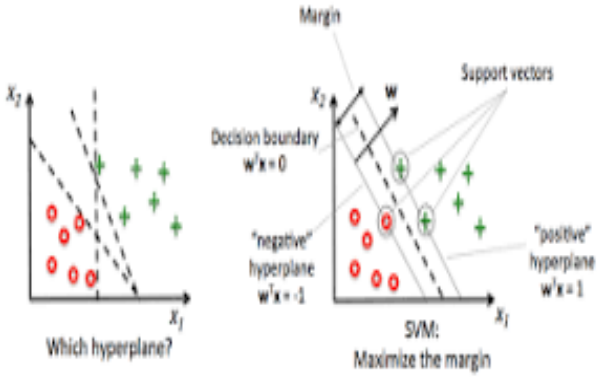

Figure: Finding the optimal hyper plane

If the classes are not linearly separable in the high dimensional space, the algorithm will add a new dimension in an attempt to further separate the classes. It will continue the process until it is able to separate the training data into two separate classes using the hyper plane.

1)Advantages of SVM

- It can deal with documents with high dimensional input space and pick out many of the irrelevant features.
- It has the ability to produce high classification accuracy compared to other text classification algorithm.

2)Disadvantages of SVM

- Training and classification time and memory requirement is comparable to decision trees, but are more expensive than naïve bayes and other algorithm.
- To choose the values of parameters in SVM is hard
- To choose the best kernel function in SVM is also a difficult problem.

Table 1:Comparison of Machine Learning techniques

Sr. No.	***Parameter Considered***	***Naïve Bayes***	***Decision Tree***	***Max Entropy***	***SVM***
1.	Parametric/Non-Parametric	Parametric	Non-Parametric	Parametric	Non-Parametric
2.	Generative/Discriminative Algorithm	Generative	Discriminative	Discriminative	Discriminative
3.	Probabilistic/Non Probabilistic	Probabilistic	Probabilistic	Probabilistic	Non-Probabilistic
4.	Performance in case of Independent/Dependent Features	Performance is good when independent features are there in the database.	Performance is good when dependent features are there in the database	Performance is good when dependent features are there in the database	Performance is good when dependent features are there in the
5.	Over fitting problem and solution	Less likely to over fit	A major disadvantage of algorithm is over fitting .It is handled by pruning of tree .	If i/p datasets are small, maxent is likely to overfit. Over fitting can be handled by using smoothing techniques,ex regularization & constraint relaxation,	To avoid over fitting SVM make use of the maximum margin hyper plane

Table 2:Comparison of recent work done with Machine Learning based Sentiment Classification

Sr.No.	**Reference No. of paper**	**Year of publication**	**Machine Learning Technique(with max accuracy/precision achieved)**	**Database Used**
1.	[14]	2011	ME,SVM(accuracy=86.66%)	Facebook data(Indonesian language)
2.	[15]	2012	NB,DT,SVM(precision=85.8%)	Twitter data(Spanish language)
3.	[16]	2013	NB,KNN,SVM accuracy >80%)	Online movie reviews
4.	[17]	2013	NB(accuracy =88.80%)	IMDB movie reviews
5.	[18]	2013	NB,SVM(accuracy =85.78%)	Digital camera reviews
6.	[19]	2014	NB,DT,KNN,SVM(accuracy =89.14%)	Online News Text data(Bangla language)
7.	[20]	2014	NB,SVM,LOGISTIC REGRESSION,BAYSIEN LOGISTIC REGRESSION,VOTED PERCEPTRON, HYPERPIPES(accuracy =92.5%)	Online political news data(Albanian language)
8.	[21]	2015	SVM(accuracy =83.1%)	Online Movie Reviews
9.	[22]	2015	NB,SVM,NB + Modified K-Means(accuracy =89.01%)	Online Mobile Reviews
10.	[23]	2016	NB,KNN,SVM(accuracy =81.71%)	Online product Reviews

VI. RESULTS AND DISCUSSION

The comparative study of Machine Learning techniques in Table 1 can be utilized in order to decide which Machine Learning technique should be used in sentiment analysis algorithm. After analyzing the recent work done in Table 2 ,it can be concluded that SVM has maximum accuracy in most cases. But since some limitations are observed with SVM ,so work needs to be done to improve accuracy and speed of overall algorithm using other Machine Learning techniques or combination of them can be used, as other Machine Learning techniques have some advantages which can be utilized and the limitations that are observed with SVM can be tackled.

VII. CONCLUSION

Sentiment Analysis is the process of extracting opinion from textual data. This survey presents the challenges faced, research areas and applications which would help the

researchers in improving the performance of automated Sentiment Analysis system.

The study of Sentiment Classification techniques gives an idea that each of the classification algorithm has its own advantages and limitations. The choice of selecting a particular classification algorithm depends on the training time required, accuracy ,memory requirements and other parameters which would ultimately help the researchers to improve overall performance of automated Sentiment Analysis system.

VIII. ACKNOWLEDGMENT

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. My sincere thanks to my guide for giving me intellectual freedom of work and guiding me time to time.

.

IX. REFERENCES

[1] Bing Liu. ,"Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, May 2012.

[2] Bo Pang and Lillian Lee , "Opinion Mining and Sentiment Analysis , Foundations and Trends_ in Information Retrieval," Vol. 2, Nos. 1–2 DOI:10.1561/1500000001, 1–135, 2008.

[3] Roberto Navigli, "Word sense disambiguation: A survey," ACM Computing Surveys Vol 41, No. 2, Article 10 , Feb 2009.

[4] David Osimo and Francesco Mureddu , "Research challenge on Opinion Mining and Sentiment Analysis," The Crossroad Roadmap on ICT for Governance and Policy Modeling 2010.

[5] A. Van Looy, "Sentiment Analysis and Opinion Mining(Business Intelligence 1) ," Social Media Management, Springer Texts in Business and Economics, DOI 10.1007/978-3-319-21990-5_7, Springer International Publishing Switzerland, 2016.

[6] A.McCallum, *and* K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification Learning for Text Categorization," Papers from the AAAI Workshop, 1998.

[7] Fabrizio Sebastiani "Machine Learning In Automated Text Categorization," ACM Computing Surveys, Vol. 34, No. 1, March 2002.

[8] J.R. Quinlan, " Induction of Decision Trees," Machine Learning 1,pp-81-106,1986.

[9] Arun K Pujari ,Data Mining Techniques, Universities Press (India) Private Limited,ISBN 8173713804,2001.

[10] B Pang B.,Lee , and L.,Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," Association for Computational Linguistics, Proceedings of the conference on Empirical Methods in Natural Language Processing, pp. 79–86,2002.

[11] Adam L. Berger Stephen A. Della Pietra ,and Vincent J. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing,"Association for Computational Linguistics Volume 22, Number 1, 1996.

[12] Vladimir N.Vapnik, "The nature of statistical Learning Theory,"Springer, New York 1995.

[13] Simon Tong and Daphne Koller, **"**Support Vector Machine Active Learning with applications to Text Classification," Journal of Machine Learning Research ,pp-45-66 2001.

[14] Aqsath Rasyid Naradhipa, and Ayu Purwarianti, **"**Sentiment classification for Indonesian message in social media," International Conference on Electrical Engineering and Informatics, Bandung, Indonesia, July 2011.

[15] Grigori Sidorov et al.," Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets," Mexican international conference on Artificial intelligence, 2012 – Springer.

[16] P.Kalaivani, Dr. K.L.Shunmuganathan ,"Sentiment classification of movie reviews by supervised Machine learning Approaches," Indian Journal of Computer Science and Engineering Vol. 4 No.4,2013.

[17] Vivek Narayanan, Ishan Arora, Arjun Bhatia " Fast and accurate sentiment classification using an enhanced Naive Bayes model," International Conference on Intelligent Data Engineering and Automated Learning ,Springer ,2013.

[18] M. Daiyan, Dr. S.K. Tiwari, and A. Alam, "To Classify Opinion of Different Domain Using Machine Learning Techniques," International Journal of Emerging Technology and Advanced Engineering , Volume 3, Issue 5, May 2013.

[19] Ashis kumar mandal and Rikta Sen , "Supervised learning methods for bangla Web document categorization," International Journal of Artificial Intelligence & Applications, Vol. 5, No. 5, September 2014.

[20] Marenglen Biba ,and Mersida Mane "Sentiment Analysis through Machine Learning: An Experimental Evaluation for Albanian," Recent Advances in Intelligent Informatics,Advances in Intelligent Systems and Computing 235 , Springer International Publishing Switzerland , 2014.

[21] Shahana P.H. and Bini Omman",Evaluation of Features on Sentimental Analysis," International Conference on Information and Communication Technologies, Procedia Computer Science 46 1585 – 1592,2015.

[22] Ashish Shukla and Rahul Misra, "Sentiment Classification and Analysis Using Modified K-Means and Naïve Bayes Algorithm," International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 8, 2015.

[23] N. Yuvaraj and A. Sabari, "Performance Analysis of Supervised Machine Learning Algorithms for Opinion Mining in E-Commerce Websites," ,Middle-East Journal of Scientific Research 24 (Techniques and Algorithms in Emerging Technologies), 2016.