ISSN No. 0976-5697

Volume 7, No. 7, Nov-Dec 2016

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

An Investigation of Efficient Machine Learning Approaches for Sentiment Classification

P.Kalarani
Research Scholar
Bharathiar University
Coimbatore, Tamilnadu, India

Dr. S.Selva Brunda
Professor & Head
Dept. of Computer Science and Engineering
Cheran College of Engineering
Karur, Tamilnadu, India

Abstract: People's opinions and experience are very valuable information in decision making process. Now-a-days several websites encourage users to express their views, suggestions and opinions related to product, services, polices, etc. publically. When a product is purchased by the customers, the process of quality evaluation is generally takes place. The interpretation of these quality evaluation results or the feelings of the consumers about the product will be helpful in determining the demand and expectations of the users towards that product. Extracting the useful content from these opinion sources becomes a challenging task. This paper reviews the machine learning-based approaches to sentiment analysis and brings out the salient features of techniques in place. The prominently used techniques and methods in machine learning-based sentiment analysis include - Naïve Bayes, Maximum Entropy and Support Vector Machine, K-nearest neighbour classification.

Keywords: opinion; sentiment analysis; machine learning; maximum entropy; support vector machines;

I. INTRODUCTION

Opinion mining also called as Sentiment analysis defines the process of determining the opinion expressed in a particular data into either positive and negative or positive, negative and neutral comments[1]. Current-day Opinion Mining and Sentiment Analysis is a field of study at the crossroad of Information Retrieval (IR) and Natural Language Processing (NLP) and share some characteristics with other disciplines such as text mining and Information Extraction. Opinion mining is a technique to detect and extract subjective information in text documents. In general, sentiment analysis tries to determine the sentiment of a writer about some aspect or the overall contextual polarity of a document.

The sentiment may be his or her judgment, mood or evaluation. A key problem in this area is sentiment classification, where a document is labeled as a positive or negative evaluation of a target object (film, book, product etc.).In recent years, the problem of "opinion mining" has seen increasing attention. Sentiment classification is a recent sub discipline of text classification which is concerned not with the topic a document is about, but with the opinion it expresses. Sentiment classification also goes under different names, among which opinion mining, sentiment analysis, sentiment extraction, or affective rating [2]. This paper will try to focus on the basic definitions of Opinion Mining, analysis of linguistic resources required for Opinion Mining, few machine learning techniques on the basis of their usage and importance for the analysis, evaluation of Sentiment classifications. The figure 1 shows the general process flow.

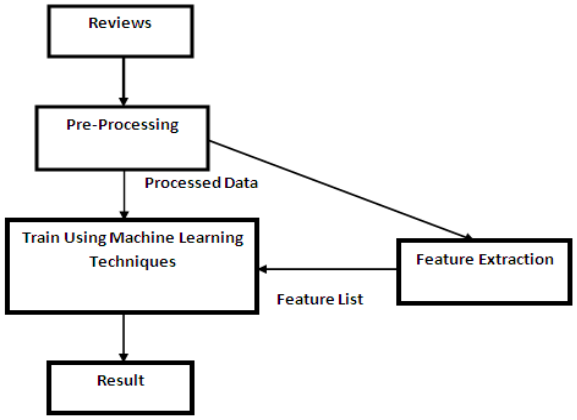

Fig.1 A general process flow using machine learning techniques

II. LITERATURE REVIEW

Bo Pang et al, [3] used machine learning techniques for sentiment analysis. The experimental setup consists of movie-review corpus with randomly selected 700 positive sentiment and 700 negative sentiment reviews. Learning methods Naïve Bayes, maximum entropy classification and support vector machines were employed.

Li, S., et al, [4] proposed semi-supervised learning for imbalanced sentiment classification. In this approach, under-sampling is performed to generate multiple sets of balanced initial training data and different random subspaces are generated dynamically which are deal with the imbalanced class distribution problem. This method provides better performance but the controlling of iteration process is challenge.

Chen, L. S., et al, [5] proposed a neural network (ANN) based method for sentiment classification in blogosphere. This approach uses semantic orientation indexes as inputs to the neural networks for determining the opinions of the bloggers rapidly and efficiently. In which, various actual blogs are used to measure the effectiveness. This approach provides better accuracy but it is not suitable for large amount of feature sets and data which increases the processing time.

Ziqiong Zhang et al, [6] proposed a method which utilizes completely prior knowledge-free supervised machine learning method. They performed sentiment analysis on written Cantonese. Their method has proved that the chosen machine learning model could be able to draw its own conclusion from the distribution of lexical elements in a piece of Cantonese review.

Moraes, R., et al, [7] proposed a document-level opinion classification approach. This approach is used to automate the process of classifying the textual review which is given on particular subject and expressing as positive or negative opinion. In this approach, SVM and ANN are effectively used as an opinion learning approach. However, in this approach, When number of terms increased SVM has less training time and high running time whereas ANN has high training time and less running time.

III. LEVELS OF SENTIMENT CLASSIFICATION

In general, sentiment analysis has been investigated mainly at three levels:

Document level: The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment. For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This task is commonly known as document-level sentiment classification. This level of analysis assumes that each document expresses opinions on a single entity (e.g., a single product). Thus, it is not applicable to documents which evaluate or compare multiple entities.

Sentence level: The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion. Neutral usually means no opinion. This level of analysis is closely related to subjectivity classification, which distinguishes sentences called objective sentences that express factual information from sentences called subjective sentences that express subjective views and opinions. However, we should note that subjectivity is not equivalent to sentiment as many objective sentences can imply opinions.

Entity or Aspect level: Both the document level and the sentence level analyses do not discover what exactly people liked and did not like. Aspect level performs finer-grained analysis. Aspect level was earlier called feature level. Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of a sentiment (positive or negative) and a target (of opinion). An opinion without its target being identified is of limited use. Realizing the importance of opinion targets also helps us understand the sentiment analysis problem better. For example, although the sentence "*although the service is not that great, I still love this restaurant*" clearly has a positive tone, we cannot say that this sentence is entirely positive. In fact, the sentence is positive about the restaurant (emphasized), but negative about its service (not emphasized). In many applications, opinion targets are described by entities and/or their different aspects. Thus, the goal of this level of analysis is to discover sentiments on entities and/or their aspects. For example, the sentence "*The iPhone's call quality is good, but its battery life is short*" evaluates two aspects, *call quality* and *battery life*, of *iPhone* (entity). The sentiment on iPhone's *call quality* is positive, but the sentiment on its *battery life* is negative. The *call quality* and *battery life* of iPhone are the opinion targets [8]. Based on this level of analysis, a structured summary of opinions about entities and their aspects can be produced, which turns unstructured text to structured data and can be used for all kinds of qualitative and quantitative analyses.

IV. MACHINE LEARNING APPROACHES

The machine learning approach belongs to supervised classification approach.This approach is more accurate because each of the classifiers is trained on a collection of representative data known as corpus. Thus, it is called "supervised learning". In a machine (supervised) learning based classification, two types of documents are required: training set and test set. A training set is used to learn the classifier and a test set is used to test the performance of the automatic classifier. Large numbers of machine learning techniques are available which classifies the opinions. Machine learning techniques like Naïve Bayes(NB), Maximum Entropy (ME) and Support Vector Machines (SVM) have achieved great success in text categorization[9].

In general, classification tasks are often divided into several sub-tasks:

1) Data preprocessing
2) Feature selection and/or feature reduction
3) Representation
4) Classification
5) Post processing

A. Naïve Bayes classification

It is an approach to text classification that assigns the class c* = argmaxc P(c | d), to a given document d. A naive Bayes classifier is a simple probabilistic classifier based on Bayes' theorem and is particularly suited when the dimensionality of the inputs are high. Its underlying probability model can be described as an "independent feature model". The Naive Bayes (NB) classifier uses the Bayes' rule Eq. (1),

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)} \qquad (1)$$

Where, P (d) plays no role in selecting c*. To estimate the term P(d|c), Naive Bayes decomposes it by assuming the fi's are conditionally independent given d's class as in Eq.(2),

$$P_{NB}(c \mid d) = \frac{P(c)\left(\prod_{i=1}^{m} P(f_i \mid c)^{n_i(d)}\right)}{P(d)} \qquad (2)$$

Where, m is the no of features and fi is the feature vector. Consider a training method consisting of a relative-frequency estimation P(c) and P (fi | c). Despite its simplicity and the fact that its conditional independence assumption clearly does not hold in real-world situations, Naive Bayes-based text categorization still tends to perform surprisingly well indeed, Naive Bayes is optimal for certain problem classes with highly dependent features[10].

B. Maximum Entropy

Maximum Entropy (ME) classification is yet another technique, which has proven effective in a number of natural language processing applications. Sometimes, it outperforms Naive Bayes at standard text classification. Its estimate Of P(c | d) takes the exponential form as in Eq. (3) [10],

$$P_{ME}(c \mid d) = \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right) \qquad (3)$$

Where, Z (d) is a normalization function. Fi, c is a feature/class function for feature fi and class c, as in Eq. (4),

$$F_{i,c}(d,c') = \begin{cases} 1 & n_i(d) > 0 \text{ and } c' = c \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

For instance, a particular feature/class function might fire if and only if the bigram "still hate" appears and the document's sentiment is hypothesized to be negative. Importantly, unlike Naive Bayes, Maximum Entropy makes no assumptions about the relationships between features and so might potentially perform better when conditional independence assumptions are not met.

C. Support Vector Machines

Support vector machines (SVMs) have been shown to be highly effective at traditional text categorization, generally outperforming Naive Bayes. They are large-margin, rather than probabilistic, classifiers, in contrast to Naive Bayes and Maximum Entropy. In the two-category case, the basic idea behind the training procedure is to find a maximum margin hyperplane [11], represented by vector w, that not only separates the document vectors in one class from those in the other, but for which the separation, or margin, is as large as possible. This corresponds to a constrained optimization problem; letting cj, {1, −1} (corresponding to positive and negative) be the correct class of document dj, the solution can be written as in Eq. (5) .

$$\vec{w} := \sum_j \alpha_j c_j \vec{d}_j, \quad \alpha_j \geq 0 \qquad (5)$$

Where, the αj's are obtained by solving a dual optimization problem. That dj such that αj is greater than zero are called support vectors, since they are the only document vectors contributing to w. Classification of test instances consists simply of determining which side of w's hyperplane they fall on.

Support vector machines were and basically attempt to find the best possible surface to separate positive and negative training samples. Support Vector Machines (SVMs) are supervised learning methods used for classification. SVM is used for sentiment classification. First module is sentiment analysis and Support vector machines perform sentiment classification task on review data. The goal of a Support Vector Machine (SVM) classifier is to find a linear hyperplane (decision boundary) that separates the data in such a way that the margin is maximized [12]. Look at a two class separation problem in two dimensions like the one illustrated in figure 1, observe that there are many possible boundary lines to separate the two classes. Each boundary has an associated margin. The rationale behind SVM's is that if we choose the one that maximizes the margin we are less likely to misclassify unknown items in the future.

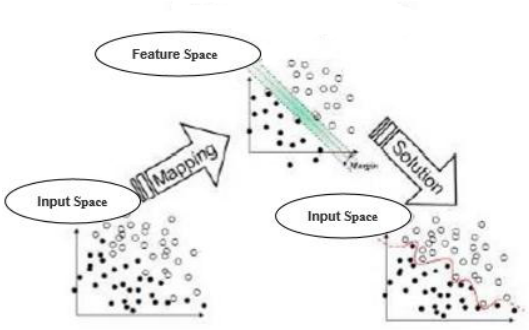

Fig.1 SVM work flow

V. STANDARD EVALUATION MEASURES

In general, the performance of sentiment classification is evaluated by using four indexes. They are Accuracy, Precision, Recall and F1-score [13]. The common way for computing these indexes is based on the confusion matrix as shown below [14]:

Table 1 Confusion Matrix

	Actual Positive	Actual Negative
Predicted Positives	True Positive	False Positive
Predicted Positives	False Negative	True Negative

These indexes can be defined by the following equations:

- $\text{Accuracy} = \frac{TN + TP}{TN + TP + FP + FN}$
- $\text{Precision} = \frac{TP}{TP + FP}$
- $\text{Recall} = \frac{TP}{TP + FN}$
- $\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Accuracy is the portion of all true predicted instances against all predicted instances. An accuracy of 100% means that the predicted instances are exactly the same as the actual instances. Precision is the portion of true positive predicted instances against all positive predicted instances. Recall is the portion of true positive predicted instances against all actual positive instances. F1 is a harmonic average of precision and recall [15].

VI. CONCLUSION

Some of the machine learning techniques like Naïve Bayes, Maximum Entropy and Support Vector Machines has been discussed. Applying Sentiment analysis to mine the large amount of unstructured data has become an important research problem. Now business organizations and individuals are putting forward their efforts to find the best system for sentiment analysis. Some of the algorithms have been used in sentiment analysis to gives good results, but no technique can resolve all the challenges. Most of the researchers reported that Support Vector Machines (SVM) has high accuracy than other algorithms, but it also has limitations. To overcome limitation of some techniques, our study focus is on the machine learning approaches for efficient sentiment classification.

REFERENCES

[1] Vikrant Hole and Mukta Takalikar, "A Survey on Sentiment Analysis and Summarization for Prediction", IJECS Volume 3 Issue 12 December, 2014 Page No.9503-9506.

[2] S. ChandraKala and C. Sindhu ISSN: 2229-6956(ONLINE) ICTACT Journal On Soft Computing, October 2012, Volume: 03, Issue: 01 "Opinion Mining And Sentiment Classification: A Survey".

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?:Sentiment classification using machine learning techniques,"inProc.ACL-02Conf.Empirical Methods Natural Lang. Process., 2002, pp. 79–86.

[4] Li, S., Wang, Z., Zhou, G., & Lee, S. Y. M. (2011, June). Semi-supervised learning for imbalanced sentiment classification. In IJCAI Proceedings-International Joint Conference on Artificial Intelligence (Vol. 22, No. 3, p. 1826).

[5] Chen, L. S., Liu, C. H., & Chiu, H. J. (2011). A neural network based approach for sentiment classification in the blogosphere. Journal of Informetrics, 5(2), 313-322.

[6] Qiang Ye, Ziqiong Zhang, Rob Law (2009),Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, Expert systems with applications, **36,** 3,pp 6527-6535.

[7] Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 40(2), 621-633.

[8] International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) Vol.4, No.1, February 2013, "Opinion Mining and Sentiment Analysis – An Assessment of Peoples' Belief: A Survey"S Padmaja and Prof. S Sameen Fatima.

[9] Geetika Gautam and Divakar Yadav, "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis", 978-1-4799-5173-4/14/$31.00 ©2014 IEEE.

[10] Rudy Prabowo, Mike Thelwall (2009), Sentiment analysis: A combined approach, Journal of Informetrics,3,2,pp 143-157.

[11] Nilesh M.Shelke, Shriniwas eshpande,Vilas Thakre(2012),Survey of techniques for opinion mining,International Journal of Computer Applications, **57,**13,pp 0975-8887.

[12] Liu, B. (2010). Sentiment Analysis and Subjectivity. To appear in Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J.Damerau), (p. 5).

[13] Rudy Prabowo, Mike Thelwall (2009), Sentiment analysis: A combined approach,Journal of Informetrics,**3**,2,pp 143-157.

[14] Xiaowen Ding, B. L. (2008). A Holistic Lexicon-Based Approach to Opinion Mining. Proceedings of the international conference on Web search and web data mining (p. 231). Palo Alto, California, USA: ACM.

[15] A. Esuli and F. Sebastiani, "SENTIWORDNET: A publicly available lexical resource for opinion mining," in Proc. 5th Conf. Lang. Res. Eval., 2006, pp. 417–422.