A Survey on Sentiment Analysis for Web-Based Data by Using Various Machine Learning Techniques

S Revathi¹, N Rajkumar²
ME, PG Scholar¹, Assistant Professor²
Department of Computer Science and Engineering
SVS College of Engineering, Coimbatore
Tamil Nadu - India

ABSTRACT
With the flourishing of social media such as blogs, forum, social network, etc... It’s an excellent source for gathering the public opinion. This paper fully focus on an exhaustively study of various machine learning techniques. Opinion mining is used to analysing and comparing the public opinion for the product review. Here we are presented various types of classification algorithm to measure sentiment analysis, semantic orientation of adjectives for some factors of subjective synonym and enrich the performance of accuracy, also shows the comparison among those classification techniques.

Keywords:- Sentiment Analysis, Semantic Orientation, Supervised Learning Algorithm, Unsupervised Learning Algorithm, NLP, WordNet

I. INTRODUCTION

Sentiment analysis is the process of understanding the emotional content of the text. In other words by using an automated tool to discern, extract and process attitudinal information found in the text. An attempt to automatically process and possibly learn from the universe of people’s online chatter. Sentiment analysis has the two types of polarity. i) Positive polarity: An object which holds the positive opinion is called the positive polarity. (e.g., awesome, joy, fun, excellent). ii) Negative polarity: an object which holds the negative opinion is called the negative polarity. (e.g., bad, worst, rubbish, terrible).

Semantic orientation is the process which focuses on the relatedness, similarity between the words. The advantage of semantic orientation of adjective is that, we cannot consider only the shortest distance to ‘excellent’ but also can consider the antonym ‘poor’[3].

A. Machine Learning Techniques

Machine learning technique is the study of computer algorithm that improves automatically through experience. This can be categories into two groups. i) Supervised Learning Method, the training data includes both the input and the desired results. Steps involved in this algorithm can be proposed in [2]. For some examples the correct results (targets) are known and are given in input to the model during the learning process. Here the training data can be manually labelled. ii) Unsupervised Learning Methods [4], is used to cluster the input data in classes on the basis of their statistical properties only. [2], [5]. The labelling can be carried out even if the labels are only available for a small number of objects representatives of the desired classes.

At the present time a variety of classification algorithms are used for sentiment analysis and to improve the performance. (e.g., [1], [2], [3], [4], [5], [6], [7],[8]). Here we present some of the classification algorithm techniques to study the sentiment classification problem.

The first kind of classification techniques are proposed in [1],[5],[6], is to examine the sentiment classification for the domain based reviews depending upon their subject matter. They used the three cross-fold validation for the three classification algorithm. In this paper focused on features based on unigrams and bigrams. Unigram presence information can be most effective and consistently better performance once unigram presence was incorporated.

The second kind of classification techniques are proposed in [2], is to classify the reviews with the four different topics based, whether it is thumbs up? (Recommended) or thumbs down? (Not recommended) by the mathematical calculation of average semantic orientation and shows their accuracy rate.

The third kind of classification technique proposed in [3], is to develop the wordNet-based measures for the semantic orientation of adjectives based on some of the factors. It includes the antonym relation as one of three strong relations between words. The comparison of the four factors shows the accuracy rate of the performance.

The fourth kind of classification technique proposed in [7] the Rule-Based Sentiment Analysis (R-BSA) is proposed to enrich the performance for the sentiment polarity for the dataset [8]. In this paper they can taken as traffic related information for the dataset. Finally compare their result with the Ku’s algorithm.

II. RELATED WORK

There has been a recent engorge of interest in routinely recognize the customer opinions, sentiment and estimation,
for the phrase. This process can be applied to many application, including classification of reviews as positive polarity or negative polarity (e.g., (Turney & Liu 2003; Dave, Lawrence, & Pennock 2003); Author Pang recognizing argumentative messages (e.g., (Sper tus 1997)), analysing the reputation of the product (e.g., (Morinaga et al. 2002; Yi et al. 2003)), tracking sentiments toward events (e.g., (Tong 2001)), genre classification (e.g., (Yu & Hatzivassiloglou 2003; Wiebe et al. 2004)), mining and summarizing reviews (Hu & Liu 2004), multi-document summarization and question answering (e.g., (Yu & Hatzivassiloglou 2003)).

Information mining systems has been developed for various domains, which includes terrorism (MUC-4 Proceedings 1992; Chieu, Ng, & Lee 2003; Riloff 1996; Soderland et al. 1995), management succession (Yangarber et al. 2000), corporate acquisitions (Freitag 1998), job postings (Califf & Mooney 1997; Freitag & McCallum 2000), rental ads (Soderland 1999; Ciravegna 2001), seminar announcements (Ciravegna 2001; Freitag & McCallum 2000), and disease outbreaks (Grishman, Huttunen, & Yangarber 2002).

Conversely, they do not conduct experiment by using the outcome of their simplification algorithm to improve the performance of an end application, stating that the performance of the algorithm is not yet satisfactory.

III. LITERATURE REVIEW

In this section various machine learning classification techniques are analysed depends upon the two major groups for the data set.

B. Naïve Bayes (NB) classifier

It is one of the statistical methods for classification, can solve problems involving both categorical and continuous valued attributes. In [1] , authors derived the Naïve Bayes classifier by observing the Bayes rule. It can be computed as,

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

First the text classification is assigned to the document level d. The class,

$$c^* = \text{arg max } P(c|d)$$ (2)

Where,

$P(d)$ plays no role in selecting the $c^*$.

The $f_i$’s are independent after decomposes of $d$’s then the class can be,

$$P_{NB}(c|d) = \frac{P(c)\prod_{j=1}^{n}P(f_i|c)^{n_i(d)}}{P(d)}$$ (3)

The calculation of relative frequency is $P(c)$ and $P(f_i|c)$ by using add-one smoothing. But this technique has the dependencies exist among variable. And it requires only small data set.

C. Maximum Entropy classification (MAXENT)

This is one of the probability distribution estimation techniques. An idea is to make fewest assumptions about the data while still being consistent with it. It is a method of data analysis. This technique has been proved an effective in a number of NLP (Natural Language Processing) [6] applications. The computation of Maximum Entropy for $P(c|d)$ is as follow as,

$$P_{ME}(c|d) = \frac{1}{Z(d)} \exp\left(\sum_i \lambda_i c F_i(c, d)\right)$$ (4)

Where,

$Z(d)$ is a normalization function,

$F_i(c, d)$ is a feature / class function for feature $f_i$ and class c.

And this can be defined as,

$$F_i(c, d) = n_i(d) > 0 \text{ and } c^{=c}$$ (5)

If the above condition (5) is true then it gives the result as one. Otherwise it'll be zero.

So that the conditional independence assumptions are not met, MaxEnt makes no assumptions about the relationships between features and so progressively perform better. It is good for large machine learning problems. (e.g. huge feature set) [1].

D. Support Vector Machine (SVM)

It is a kernel based method. It is based on the training data set. It is mainly used to find the similarity of the text. For an example the word ‘good’ has the positive polarity. ‘Excellent’ also has the positive polarity. These two words are used to find the similarity among them. This method is one of the highly an effectual categorization. The corresponding search is used to a embarrassed optimization problem. Let $c_j \in \{-1, 1\}$ be the document class of $d_j$. The formula can be written as,

$$\vec{w} = \sum_j a_j \alpha_j c_j d_j, \alpha_j \geq 0$$ (6)

Where,

$\vec{w}$ is represented as a hyper plane vector.

The $a_j$’s are obtained by solving a dual optimization problem. $d_j$ such that $a_j$ is always greater than zero called support vector.[8]

The main advantages of SVM is that maximization of generalization ability, it has no local minima, and robustness to outliers.

E. Unsupervised Learning Techniques(ULT)
If the data has to be processed by using machine learning methods, where the desired output is not given then the learning task is called unsupervised learning method [2],[4],[5]. This method involved the steps; First step is to extract phrases containing adjectives or adverbs. The second step is to estimate the semantic orientation of each phrase. The final step is to calculate the average semantic orientation.

This algorithm is only used to extract the phrases but in [2] the past work established that the adjective in the phrases are the good indicators of the subjective.

**F. Pointwise Mutual Information and Information Retrieval (PMI-JR) Algorithm**

In [2] an author Peter D.Turney proposed this algorithm is an essentially to calculate the semantic orientation of the similarity of the words .For an Example the word ‘excellent’ is compare with ‘poor’.PMI between two words can be defined as,

\[ \text{PMI}(\text{word}_1, \text{word}_2) = \log_2 \left( \frac{p(\text{word}_1 \& \text{word}_2)}{p(\text{word}_1)p(\text{word}_2)} \right) \]

where,

\[ p(\text{word}_1 \& \text{word}_2) \] is the probability of the co-occurrence of the \text{word}_1 and \text{word}_2

\[ p(\text{word}_1)p(\text{word}_2) \] is the probability of co-occur.

The ratio between \[ p(\text{word}_1 \& \text{word}_2) \] and \[ p(\text{word}_1)p(\text{word}_2) \] is a measure of the degree of statistical dependence between the words. The log ratio is the amount of information that presence one of the words. So here the semantic orientation (SO) is calculated for a phrase is as follows as,

\[ \text{SO (phrase)} = \text{PMI (phrase, “excellent”) – PMI (phrase, “poor”)} \]

From here if the phrase is strongly coupled with ‘excellent’ then it is called as positive. Else if the phrase is strongly coupled with ‘poor’ then it is called as negative. But in [2] Turney,2001 projected that the previous work has shown that NEAR operator limitation performs better than AND while measuring the strength of semantic association between words.

**G. Linear Classifier Technique (LCT)**

Frequently this technique is applied in the statistical technique recurrently. It is used to describe the relationship between one variable and the values taken by several other variables [5]. The commonly used form of regression model is the general linear model formally it can be written as,

\[ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n + \epsilon \]

By applying this equation to each of the given samples a new set of equations can be obtained,

\[ y_j = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + ... + \beta_n x_{nj} + \epsilon_j \]

where, \( \epsilon_j \)'s are called regression errors for each \( m \). This technique is called as linear because expected value of \( y_j \) is a linear function.

**H. Ku’s Algorithm**

In [7], author compared their result with this Ku’s algorithm. First this algorithm detects the sentiment words. Second step is to identify the opinion of the sentences and as a final point documents. This algorithm is written for both sentence level and document level for the review. To estimate the inclination between the sentence level and the document level, then the sentiment words are occupied. To detect the sentiment polarity of the phrase in Chinese documents, then they require a Chinese dictionary. Nevertheless, a small dictionary may not enough for the problem of coverage.

They expand a method to learn sentiment words and their strengths from manifold resources. First they are collect two sets of sentiment words, including General Inquirer (abbreviated as GI) and Chinese Network Sentiment Dictionary (abbreviated as CNSD). The former is in English and we translate those words into Chinese. The latter, whose sentiment words are collected from the Internet, is in Chinese. Words from these two resources form the “seed vocabulary” in our dictionary.

**I. Rule-Based Sentiment Analysis algorithm(R-BSA)**

In [7], a new algorithm is proposed called Rule-Based sentiment Analysis algorithm. It governs an association of class of rule mining algorithm, is to mechanically determine an interesting and effective rules capable of extracting product features, sentence opinion for a specific product. This algorithm is used to take the description of the product feature. (e.g., price, design, battery, etc.), all of the input must be provided by the human and it incorporates with the World Wide Web (WWW). Also a set of seed opinion words, a lexical dictionary is used to automatically extract the opinion of a sentence. It was provided the higher accuracy rates when compared to the Ku’s algorithm.

**IV. COMPARISON OF MACHINE LEARNING TECHNIQUES**

For any kind of review, there must be a various learning techniques are used to sentiment analysis, semantic orientation of adjectives and the performance in an accuracy rate. The main aim is to analysis the sentiment polarity, and then accuracy rate. Here Table I provided the comparison among the machine learning algorithms.
From this table in [1],[6], the three cross fold validation can be calculated for the review. The main aim of this paper is to examine the effectiveness of applying machine learning techniques to sentiment classification problem. But here they can consider only the sentiment polarity not the semantic orientation of adjectives.

In [2],[4],[5],[6], objective of this paper is to classify the reviews by using the learning algorithm and produces as output as recommended (thumbs up) or not recommended (not recommended).They can compare their product with another branded product from customer point of view. They can consider the four domain specific review as their input dataset. Also performance of accuracy rate is better accepted.

In [3], the approach is used to measures the semantic orientation of adjectives based on WordNet. Here they consider the four different factors and evaluate the intersection of words for the accuracy rate.

In [7], the main aim is to classifying the data from the training set. The traffic related information is taken as an input review. In this paper the accuracy rate is compared with Ku’s algorithm.

<table>
<thead>
<tr>
<th>S. No</th>
<th>Machine Learning Algorithm</th>
<th>Protocol</th>
<th>Goals</th>
<th>Gain</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>NB, SVM, MAXENT</td>
<td>Find the three cross fold validation</td>
<td>Best performance in accuracies</td>
<td>Unreliable for low frequency feature</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>ULT, PMI-IR</td>
<td>Determine the semantic orientation of adjectives</td>
<td>Domain specific</td>
<td>For large dataset, distant search engine for rare words</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>LCT</td>
<td>Opinion mining</td>
<td>Automatically identify the product feature</td>
<td>Strength of opinions and accuracy rate</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>KU’S R-BSA</td>
<td>Addressed the key problems of TSA, including the design of architecture and the construction of related bases</td>
<td>R-BSA provided the higher accuracy rate</td>
<td>Difficult to classify the low importance sentences.</td>
<td></td>
</tr>
</tbody>
</table>

V. EVALUATION RESULTS

All of the learning techniques had been provide the accuracy rate for the various sample dataset in this survey. In [1], Naïve Bayes (NB) classifier it was provided the accuracy rates of 78.7%. But this has the dependencies exist among variables. Maximum Entropy (MAXENT) provided the accuracy rate of 77.7%. This is very useful for large machine learning problems (e.g., huge feature set). Support Vector Machine (SVM), was provide the accuracy rate of 81.9%. While comparing this three cross fold validation the SVM was provided the best accuracy rate result.

In [2], the proposed algorithm classifies adjectives with accuracies ranging from 78% to 92%. But this result was depending on the amount of training dataset that is available.

In [3], the proposed technique can be used to estimate the four factors. First factor, Evaluative (EVA) was provided the accuracy result of 68.19% for 349 words. Second factor, Potency (POT) was provided the accuracy rate of 71.36% for 419 words. Third factor, Activity (ACT) was provided the accuracy result of 61.85% for 173 words. The fourth factor, Evaluative II (EVA) was provided the accuracy rate of 67.32% for 667 words.

In [7] the proposed R-BSA algorithm was provided the accuracy result of 82.45% when compared with Ku’s algorithm, it provided only the 65.85% of accuracy result. So in this paper their R-BSA was provided the best accuracy result.

Fig.1 shows the evaluation result for various machine Learning Techniques accuracy rate in percentage. Here we can considered only the Naive Bayes (NB), Maximum Entropy (MAXENT), Support Vector Machine (SVM), KU,
Rule-Based Sentiment Analysis (R-BSA) protocols for the evaluation.

VI. CONCLUSION

In this survey we have focused the sentiment analysis, semantic orientation of adjectives and performance of it’s accuracy rate. Some of the proposed techniques and some common drawbacks are identified. But still low level frequency sentence cannot be performed for the sentiment polarity. The idea of sentiment analysis should be further enriched the stylistic feature of the text and also enrich the performance of accuracy rate by using MapReduce concept of hadoop in big data environment.

REFERENCES

[5] Comparison of Classification Methods Based on the Type of Attributes and Sample Size Reza Entezari-Maleki*Corresponding author, Arash Rezaei, and Behrouz Minaei-Bidgoli.
[6] A Multi-Class SVM Classification System Based on Methods of Self-Learning and Error Filtering JuiHis Fu ChihHsiung Huang SingLing Lee.
[7] Jianping Cao, Ke Zeng, Hui Wang, Member, IEEE, Jiajun Cheng, Fengcai Qiao, Ding Wen, Senior Member, IEEE, and Yanqing Gao, Member, IEEE,” Web-Based Traffic Sentiment Analysis: Methods and Applications”.