Journal of Theoretical and Applied Information Technology
20th January 2014. Vol. 59 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

INVESTIGATION OF SUPPORT VECTOR MACHINE CLASSIFIER FOR OPINION MINING

[1]K.SARASWATHI, [2] Dr. A.TAMILARASI

[1]KONGU ENGINEERING COLLEGE, Department Of Computer Technology, ERODE, TAMILNADU, INDIA

[2]KONGU ENGINEERING COLLEGE, Department Of Computer Applications, ERODE, TAMILNADU, INDIA

E-mail: [1] paran_2013@rediffmail.com, [2] drtaav@rediffmail.com

ABSTRACT

Complicated text understanding technology which extracts opinion, and sentiment analysis is called opinion mining. Building systems to collect/examine opinions about a product in blog posts, comments, and reviews/tweets is sentiment analysis. Product reviews are the focus of existing work on review mining and summarization. This study focuses on movie reviews, investigating opinion classification of online movie reviews based on opinion/corpus words used regularly in reviewed documents.

Keywords: *Opinion Mining, Classification Accuracy, Sentiment analysis, Movie reviews, Support Vector Machine, Bagging*

1. INTRODUCTION

Opinion mining [1] is a sub-discipline of computational linguistics extracting people's opinions from the web. Web expansion encourages users to contribute and express themselves through blogs, videos and social networking sites, all of which generate phenomenal amount of information that needs to be analyzed. In a set of evaluative text documents, *D having* an object's opinions [2] (or sentiments), opinion mining plans to extract that object's attributes and components commented on in each document $d \in D$ and detect whether comments are positive/negative or neutral.

Sentiment analysis [3] tracks public mood, through a type of natural language processing about a specific product/topic. Sentiment analysis, also called opinion mining, includes building systems to collect/examine a product's opinions from various media outlets on the net. Sentiment analysis is useful in many ways. In marketing, for example, it aids judging an ad campaign or new product launch successfully, determines what product versions or service are popular even identifying demographics which like/dislike specific features. Literature survey indicates two techniques including machine learning and semantic orientation.

Opinions are expressed on anything, e.g., product, service, topic, an individual, organization, or event. The term *object* denotes the entity commented upon. An object has a components set, and an attributes set. Each component has its sub-components and attributes set etc. So an object based on the *part-of* relationship can be hierarchically decomposed. An opinion's *semantic orientation* on a feature *f reveals* whether it is positive, negative or neutral. A *model* for an object and opinions set on its features is defined as a *feature-based opinion mining model*.

Opinion mining and sentiment analysis have many applications.

- Argument mapping software policy statements are organized logically by explicating logical links in them. Online Deliberation tools like Compendium, Debatepedia, Cohere, Debategraph were developed to provide a logical structure to many policy statements, thereby linking arguments with their back up evidence.
- Voting Advise Applications help voters find out which political party (other voters) has positions closer to theirs.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Automated content analysis processes qualitative data. Today there are many tools combining statistical algorithm with semantics and ontologies, as also machine learning with human supervision, all of which identify relevant comments, assigning positive or negative connotations [1].

Machine learning sentiment analysis usually comes under supervised classification and under text classification techniques in specific. Two sets of documents; training and test set are required in machine learning classification. Automatic classifiers use training set to learn a document's differentiating characteristics, whereas a test set validates automatic classifier performance.

Semantic orientation in Sentiment analysis is unsupervised learning, as it needs no earlier training to mine data. It just measures how positive or negative a word might be.

Sentiment classification [4] is generally two-class classification *positive* and *negative* problem. Training /testing data are usually product reviews. As online reviews reviewer assigned rating scores e.g., 1-5 stars, ratings decide positive and negative classes. For example, a 4 or 5 star review is considered a positive review while that with 1 to 2 stars is thought a negative review. Research papers do not use neutral class as this ensures easier classification. However, it is possible to use neutral class by assigning 3-stars in reviews.

Sentiment classification is basically text classification. Traditional text classification classifies different topic documents like politics, sciences, and sports where topic related words become key features. But in sentiment classification, sentiment/opinion words indicating positive/negative opinions are more important. Examples are *great*, *excellent*, *amazing*, *horrible*, *bad*, *worst*, etc. As sentiment words dominate sentiment classification, it goes without saying that sentiment words/phrases might be used in an unsupervised way [5]. Classification here is based on fixed syntactic patterns - composed based on part-of-speech (POS) tags - which express opinions.

This study investigates online movie review opinion classification based on opinion words/corpus words used regularly in documents being reviewed. Feature set from reviews is extracted through the use of Inverse document frequency with reviews being classified as positive or negative by using Support Vector Machine. The section which follows briefly reviews related works in literature, describes materials, methods, classification algorithms, describes results and finally discusses the same.

2. RELATED WORKS

A unified collocation framework (UCF) was proposed by Xia, et al., [6] which described a unified collocation-driven (UCD) opinion mining procedure. UCF incorporates attribute-sentiment collocations and its syntactical features for achieving generalization ability. Early experiments revealed that 0.245 on average improved opinion extraction recall without losing opinion extraction precision and accuracy in sentiment analysis.

Opinion mining extracts opinions by a source on a specific target, from a document set. A comparative study on methods/resources used for opinion mining from newspaper quotations was presented by Balahur, et al., [7]. Annotated quotations from news evaluated the proposed approaches using EMM news gathering engine. A generic opinion mining system uses big lexicons and also specialized training/ testing data.

A novel approach for mining opinions from product reviews was proposed by Wu, et al., [8], where opinion mining tasks were converted to identify product features, opinion expressions and their inter relations. A concept of phrase dependency parsing which took advantage of the product features being phrases was introduced. This concept extracted relations between product features and opinion expressions. Evaluations showed that mining tasks benefited from this.

Plantie, et al., [9] classified documents according to their opinions and value judgments. The originality of the proposed approach combined linguistic pre-processing, classification and a voting system through many classification procedures. Document representation determined features to store textual data in data warehouses. Experiments from a text mining French challenge corpora (DEFT) showed the approach to be efficient.

Opinion mining identifies whether expressed opinion on a topic in a document is positive or negative. Saleh, et al., [10] explored this using Support Vector Machines (SVM) to test various data set domains through the use of many weighting schemes. Experiments were undertaken with varied features on three corpora, two of which had been already used in many works. The last one was built from Amazon.com to prove SVM feasibility in different domains.

Journal of Theoretical and Applied Information Technology
20th January 2014. Vol. 59 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Maynard, et al., [11] discussed opinion mining related issues from social media and their challenges on a Natural Language Processing (NLP) system. This was accompanied by 2 example applications developed in various domains. In contrast to machine learning techniques related to opinion mining work, the new system engendered a modular rule-based approach to perform shallow linguistic analysis. It builds on many linguistic subcomponents to generate final opinion, on polarity and score.

Pak, et al., [12] used popular microblogging platform Twitter for sentiment analysis, which revealed how to collect a corpus for sentiment analysis and opinion mining task automatically. The system performs the collected corpus's linguistic analysis and explained discovered phenomena. It was able to build a sentiment classifier capable of determining a document's positive, negative and neutral sentiments. Evaluations proved the efficiency of the proposed techniques as they performed better than earlier methods.

3. MATERIALS AND METHODS

3.1 Dataset

Pang and Lee [13] movie reviews data set containing 2,000 movie reviews with 1,000 positive and 1,000 negatives evaluated classification algorithms. An earlier version with 700 positive and 700 negative reviews was also used in Pang, et al., [14]. Positive/negative classification as specified by the reviewer is extracted from ratings automatically. The dataset included only reviews whose rating was indicated by stars or a numerical system. This study uses a subset of 150 positive and 150 negative opinions.

3.2 Feature Extraction

Features are extracted using Inverse Document Frequency (IDF) for document classification. Also prepared was a list of stop words (commonly occurring words) and stemming words (words with similar context). The terms document frequency (df) which includes a number of documents having the term is computed. Rarely occurring terms are more informative than those which occur frequently. Thus, rare words are assigned higher weights than those used regularly. Captured by document frequency term t (df_t), inverse document frequency (idf_t) represents scaling factor [15]. Term t's importance is scaled down when used frequently. The idf_t is defined as follows:

$$IDF(a) = \log \frac{1+|x|}{x_a}$$

x_a is the set of documents containing the term a.

3.3 Classifier

SVM classification has roots in structural risk minimization (SRM) that determines classification decision function through empirical risk minimizing [16]

$$R = \frac{1}{l} \sum_{i=1}^{L} | f(X_i) - y_i |,$$

where L and f are examples size and classification decision function, respectively. Determining optimal separating hyperplane which ensures low generalization error is of primary concern for SVM. Classification decision function in a linear separable problem is represented by

$$f_{w,b} = sign(w \cdot x + b).$$

Optimal separating hyperplane in SVM is determined through largest separation margin between classes bisecting the shortest line between two class's convex hulls. Optimal hyperplane satisfies constrained minimization as

$$Min \ \frac{1}{2} w^T w,$$
$$y_i (w \cdot x_i + b) \geq 1.$$

SVM methods are used routinely for classification. For specific training data (x_i, y_i), i=1,...n, where $x_i \in \Re^d$ is a feature vector and $y_i \in \{+1, -1\}$ indicates class value to solve the following optimization problem:

$$w = \sum_{i=1}^{n} \alpha_i y_i \Phi(x_i)$$

x_i is a support vector if $\alpha_i \neq 0$ New instance x is computed by the following function:

$$f(x) = \sum_{i=1}^{n_s} \alpha_i y_i K(s_i, x) + b$$

Where s_i are support vectors and n_S number of vectors and polynomial kernel function is given by:

Journal of Theoretical and Applied Information Technology
20th January 2014. Vol. 59 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \quad \text{where } \gamma > 0$$

And the Radial basis function (RBF) kernel:

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \quad \text{where } \gamma > 0$$

Sequential Minimal Optimization (SMO) [17] are algorithms that quickly solve SVM QP problem, expand QP without extra matrix storage and do not take recourse to numerical QP optimization. SMO's advantage is the ability to solve Lagrange multipliers analytically. SMO is a supervised learning algorithm for classification/regression for quick SVM implementation. Its advantage is its attempts to maximize margins, the distance, for example, between classifier and nearest training datum. SMO constructs a hyperplane/hyperplanes set in n-dimensional space for classification. When a hyperplane has large distance to nearest training data class points a separation is considered good. Generally, larger the margin, lower the classifier generalization error.

3.4 Bagging

Bagging improves classification and regression trees stability and predictive power [18]. It is a general technique applicable in various settings to improve predictions and hence its use is not restricted to improving tree-based predictions alone. Breiman shows how bagging improves predictions and performance variability when data sets are considered [19]. Bagging reduced CART's misclassification rates by 6% to 77% when classification examples were examined.

The problem of predicting a numerical response variable's value, Yx, resulting from or occurring with a given set of inputs, x, should be considered to understand how and why bagging works and determines situations where bagging can induce improvements. φ(x), is a prediction from using a particular process like CART, or OLS regression through the use of a specified method for model selection. Allowing μφ denote E (φ(x)), where expectation regarding distribution underlying the learning sample (viewed as a random variable, φ(x) is a learning sample function seen as a high-dimensional random variable) and not x (considered fixed), the following equations result.

$$\left(E\left[y_x - \mu_\varphi\right]^2\right) = \left(E\left[y_x - \mu_\varphi\right] + \left[\mu_\varphi - \varphi(x)\right]^2\right)$$
$$= \left(E\left[y_x - \mu_\varphi\right]^2\right) + 2E\left(y_x - \mu_\varphi\right)E\left(\mu_\varphi - \varphi(x)\right) + E\left[\mu_\varphi - \varphi(x)\right]^2$$
$$= \left(E\left[y_x - \mu_\varphi\right]^2 + E\left[\mu_\varphi - \varphi(x)\right]^2\right)$$
$$= \left(E\left[y_x - \mu_\varphi\right]^2\right) + \text{var}\,ience\left(\varphi(x)\right)$$
$$\geq \left(E\left[y_x - \mu_\varphi\right]^2\right)$$

The future response independence Yx, and learning sample based predictor φ(x), is used. Predictor variance φ(x) is positive (as all random samples do not yield prediction sample value), as in nontrivial situations to ensure strict inequality leading to the result that if μφ = E(φ(x)) is a predictor, it would lower mean squared prediction error than does φ(x).

4. RESULTS AND DISCUSSION

An Internet Movie Database (IMBd) subset having 300 instances (150 positive and 150 negative) classified by the new method is used for evaluation. The following tables and figures provide classification accuracy, Root mean squared error (RMSE), precision and recall for SVM for classifying opinions as either positive or negative.

Table 1: Classification Accuracy And RNSE For Various Classifiers Used

Technique used	**Classification Accuracy**	**RMSE**
SVM with Polykernel	87.00%	0.3606
SVM with RBF Kernel	73.33%	0.5164
Bagging with SVM	**88.00%**	**0.2836**

Journal of Theoretical and Applied Information Technology
20th January 2014. Vol. 59 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

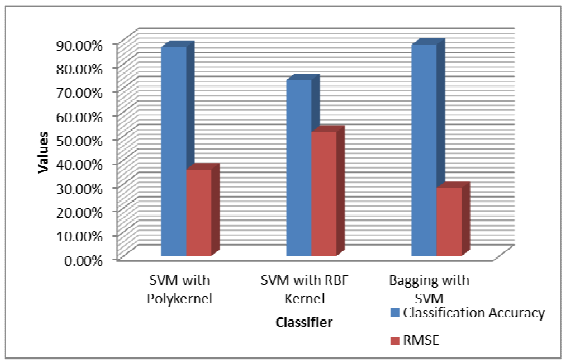

Figure 1: Classification Accuracy and RMSE for various classifiers used

It is seen from Figure 1, that the classification accuracy achieved by Bagging with SVM is much better than SVM Polykernel or RBF. The RMSE is also less for bagging with SVM. The precision, recall and f Measure values are given by:

$$\text{Pr}ecision = \frac{True \text{ positives}}{True \text{ positives} + false \text{ positives}}$$

$$\text{Re}\,call = \frac{True \text{ positives}}{True \text{ positives} + false \text{ negatives}}$$

$$fMeasure = 2 * \frac{precision * recall}{precision + recall}$$

Table 2: Precision and Recall values

Technique used	Precision	Recall	F Measure
SVM with Polykernel	0.87	0.87	0.87
SVM with RBF Kernel	0.76	0.733	0.726
Bagging with SVM	0.881	0.88	0.88

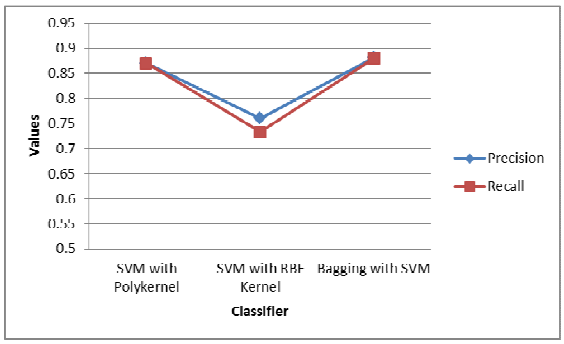

Figure 2: Precision and Recall

It is observed from Figure 2 that the precision and recall of Bagging with SVM. As the recall is also high, most relevant results are returned.

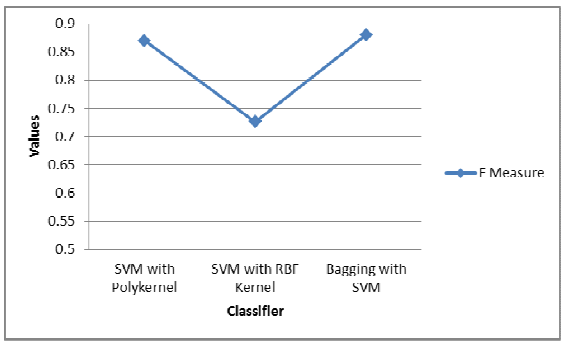

Figure 3: f Measure

5. CONCLUSION

This study uses SVM to classify as positive or negative feature sets from reviews extracted through the use of Inverse document frequency. SVM classifies features using Polykernel, RBF kernel. They are also classified using Bagging with SVM. A subset of Internet Movie Database (IMBd) with 300 instances (150 positive and 150 negative) was used for evaluation.

REFRENCES

[1]. Osimo, D., and Mureddu, F., Research Challenge on Opinion Mining and Sentiment Analysis

[2]. Liu, B., Opinion Mining.

[3]. Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment Analysis and Opinion Mining: A Survey. *International Journal*, *2*(6).

[4]. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1-167.

[5]. Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics.

[6]. Xia, Y. Q., Xu, R. F., Wong, K. F., & Zheng, F. (2007, August). The unified collocation framework for opinion mining. In *Machine Learning and Cybernetics, 2007 International Conference on* (Vol. 2, pp. 844-850). IEEE.

[7]. Balahur, A., Steinberger, R., Goot, E. V. D., Pouliquen, B., & Kabadjov, M. (2009, September). Opinion mining on newspaper quotations. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on* (Vol. 3, pp. 523-526). IET.

[8]. Wu, Y., Zhang, Q., Huang, X., & Wu, L. (2009, August). Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3* (pp. 1533-1541). Association for Computational Linguistics.

[9]. Plantié, M., Roche, M., Dray, G., & Poncelet, P. (2008). Is a voting approach accurate for opinion mining?. In *Data Warehousing and Knowledge Discovery* (pp. 413-422). Springer Berlin Heidelberg.

[10]. Rushdi Saleh, M., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. (2011). Experiments with SVM to classify opinions in different domains.*Expert Systems with Applications*, *38*(12), 14799-14804.

[11]. Maynard, D., Bontcheva, K., & Rout, D. (2012). Challenges in developing opinion mining tools for social media. *Proceedings of@ NLP can u tag# user_generated_content*.

[12]. Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC* (Vol. 2010).

[13]. Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of ACL 2004, pp.271-278.

[14]. Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of EMNLP 2002, pp.79-86.

[15]. Papineni, K. (2001, June). Why inverse document frequency? In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (pp. 1-8). Association for Computational Linguistics.

[16]. Kim, H. C., Pang, S., Je, H. M., Kim, D., & Yang Bang, S. (2003). Constructing support vector machine ensemble. Pattern recognition, 36(12), 2757-2767.

[17]. Alazab, M., Venkatraman, S., Watters, P., & Alazab, M. (2011). Zero-day malware detection based on supervised learning algorithms of API call signatures. In *AusDM 11: Proceedings of the Ninth Australasian Data Mining Conference* (pp. 171-182). Australian Computer Society.

[18]. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont

[19]. Breiman, L.: Bagging predictors. Machine Learning 24, 123–140 (1996).