Journal of Theoretical and Applied Information Technology
10th December 2013. Vol. 58 No.1

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

OPINION MINING USING DECISION TREE BASED FEATURE SELECTION THROUGH MANHATTAN HIERARCHICAL CLUSTER MEASURE

JEEVANANDAM JOTHEESWARAN[1], DR. Y. S. KUMARASWAMY[2]
[1]Research Scholar, Vel Tech Dr. RR & Dr. SR technical University, Chennai, INDIA.
Asst.Professor, CSE Department, HKBK College of Engineering, Bangalore, INDIA.
[2]HOD & Sr.Prof, Dept. of MCA, Dayananda Sagar College of Engineering, Bangalore, INDIA.
Email [1]: jsearch@zmail.com [2]ysklds wamy2@yahoo.com

ABSTRACT

Opinion mining plays a major role in text mining applications in consumer attitude detection, brand and product positioning, customer relationship management, and market research. These applications led to a new generation of companies and products meant for online market perception, reputation management and online content monitoring. Subjectivity and sentiment analysis focus on private states automatic identification like beliefs, opinions, sentiments, evaluations, emotions and natural language speculations. Subjectivity classification labels data as either subjective or objective, whereas sentiment classification adds additional granularity through further classification of subjective data as positive/negative or neutral. Features are extracted from the data for classifying the sentiment. Feature selection has gained importance due to its contribution to save classification cost with regard to time and computation load. In this paper, the main focus is on feature selection for Opinion mining using decision tree based feature selection. The proposed method is evaluated using IMDb data set, and is compared with Principal Component Analysis (PCA). The experimental results show that the proposed feature selection method is promising.

Keywords: *Opinion Mining, Imdb, Inverse Document Frequency (IDF), Principal Component Analysis (PCA), Leaningr Vector Quantization(LVQ).*

1. INTRODUCTION

A text understanding technology, Opinion mining assists people locate relevant opinions in a large review collection volume. An opinion mining technology based search engine shows potential to address this issue. An opinion-mining tool pores over product reviews for extraction of opinion units saving them in opinion databases. When users input opinion-searching query, search engine extracts product names and attribute from query, forwards a complicated SQL query to opinion database and displays output on Web interface. Opinion-search engines arrange information based on opinion and not on the document. Hence, product review information is accessed quickly/easily [1].

State-of-the-art opinion mining techniques are divided into 2 camps, i.e. attribute-driven methods and sentiment-driven methods. Their basic idea is to use either attribute or sentiment keyword to locate opinion candidates through application of certain opinion patterns (involving attributes/sentiment keywords) for extraction of sentiment expressions filtering false opinion candidates. A drawback with this method is that they yield higher *precision* at the cost of large *recall* loss as generalization ability is not implied. The problem is mainly caused by out-of-vocabulary (OOV) attributes and OOV sentiment keywords being encountered in natural language review text.

Sentiment analysis is a natural language processing type to track public mood about a specific product or a topic. Sentiment analysis, also called opinion mining, involves building a system for collecting and examining opinions about a product in comments, blog posts, tweets or reviews. Sentiment analysis is used in many ways. For example, it judges the success of an ad campaign/new product launch in marketing to determine which product versions or service are popular and identify which demographics like/dislike a specific features [2].

There are many challenges to Sentiment analysis. The first is an opinion word considered positive in one situation and negative in another. The second challenge is that people express opinions in various ways. Conventional text

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

processing is based on the fact that limited differences can be identified between two text pieces which does not change meaning much.

Some research fields are predominant in Sentiment analysis: sentiment classification, feature based Sentiment classification and opinion summarization. Sentiment classification classifies whole documents according to opinions to specific objects. But feature-based Sentiment classification considers certain subjects features opinions. Opinion summarization is different from traditional text summarization as the only product features are mined on which customers expressed opinions. Opinion summarization fails to summarize reviews by choosing a subset or rewrites some original sentences from reviews to capture main points as in traditional text summarization.

It is hard for a human reader to locate relevant sources, extract related sentences and opinions, read, summarize, and organize them into usable forms. Thus, automated opinion discovery or summarization systems are needed. *Sentiment analysis*, also called *opinion mining*, came from this need and is a challenging natural language processing/text mining problem. It's huge value for applications led to its explosive growth in research, academia and industry. It focuses on the topics below [3]:

The problem of sentiment analysis: A scientific problem has to be defined before it is solved to formalize it. Formulation introduces basic definitions, core concepts/issues, sub-problems and target objectives. It is also a framework to unite different research directions. From an application point of view, it tells practitioners what are the main tasks, inputs and outputs and how resulting outputs are used in practice.

Feature-based sentiment analysis: This discovers targets on which opinions were expressed in a sentence, and determines whether opinions are positive/negative or neutral. The targets are objects, and their components/tributes/features. An object could be a service, product, organization, individual, topic, event etc. For example, a product review sentence identifies product features commented on by reviewer determining whether comments are positive/negative.

Frequently used data mining dimensionality reduction technique is a feature selection that selects an original features subset based on specific criteria. It reduces features number, removes irrelevant/redundant/noisy data, providing applications effects which include speeding up data mining algorithms, improving mining performance like predictive accuracy and result comprehensibility. Feature selection is an active research field and developed machine learning, and data mining for years and is now applied to fields like text mining, genomic analysis, intrusion detection and image retrieval. When new applications emerged, many challenges also arose needing new theories/methods to address high-dimensional/complex data. Optimal redundancy removal, stable feature selection, and auxiliary data and prior knowledge exploitation in feature selection are among the fundamental and challenging problems in feature selection. Up-to-date, large volumes of literature were published on the research direction of feature selection.

Inverse document frequency (IDF) is an important and widely used concept in information retrieval. When IDF combines with term frequency (TF), it results in a robust/highly effective term weighting scheme applied across various application areas like databases natural language processing, knowledge management, text classification and information retrieval. There were few attempts to improve limited number of "classical" IDF formulations mainly due to the fact that it is nontrivial to change standard IDF formulation in a theoretically meaningful way while improving effectiveness. There may be heuristic ways to alter IDF formulation, but doing so leads to little understanding as to why things improved.

In this paper, it is proposed to compute the inverse document frequency and select features using proposed feature selection and compare it with Principal Component analysis. The effectiveness of the features thus selected is evaluated using LVQ classifier. It is proposed to extract the feature set from IMDb movie data set.

2. RELATED WORK

Online customer reviews are a significant informative resource useful for both potential customers and product manufacturers. Reviews are written in natural language and are unstructured-free-texts scheme in web pages. The task of manually scanning huge amounts of reviews is computationally burdensome and not practically implemented regarding businesses/customer perspectives. Hence, it is

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

efficient to automatically process various reviews providing necessary information in a correct method. Opinion summarization addresses how to determine sentiment, attitude/opinion an author expressed in natural language text regarding a specific feature. An approach to mine the product feature and opinion based on both syntactic and semantic information considerations was proposed by Somprasertsri and Lalitrojwong [4]. Application of dependency relations and ontological knowledge with probabilistic based model, proved that this method was more flexible than others.

Opinion mining extracts tasks from documents opinions as expressed by sources on a target. A comparative study on methods used for mining opinions from the newspaper article quotations. Its difficulty in being motivated by various possible targets and variety that quotes have, was presented by Balahur, et al., [5]. This approach evaluated annotated quotations from news provided from the EMM news engine. Generic opinion mining requires the use of large lexicons, and specialized training/testing data.

In the past, researchers developed large feature selection algorithms designed for other purposes and each model had its own advantages/disadvantages. Though there were efforts to survey existing feature selection algorithms, a repository collecting representative feature selection algorithms to facilitate comparison/joint study is yet to materialize. To offset this, Zhao, et al., [6] presented a feature selection repository designed to collect popular algorithms developed in feature selection research to be a platform to facilitate application/ comparison/joint study. The repository assists researchers achieve reliable evaluation when developing the new feature selection algorithms.

In schools/colleges, student comments about the courses are an informative resource to improve teaching effectiveness. El-Halees and Gaza [7] proposed a model to extract students' opinions knowledge to improve and measure course performance. The task is to use student generated contents to study specific course's performance and to compare it with that of other courses. A model was suggested for this consisting of 2 components: Feature extraction to extract features like teachers, exams and resources from user-generated content for a selected course and classifier to provide sentiment for each feature. Then they are grouped and features visualized graphically. This ensures comparison of one or more courses.

Faster and accessible internet ensures that people search/learn from fragmented knowledge. Generally, huge volumes of documents and homepages or learning objects are returned by search engines without any specific order. Even if related, a user moves forward/backward in the material to figure out the page to be read first as users usually have little or no experience in that domain. Though a user may have domain intuition they are still to be linked. A learning path construction approach based on modified TF-IDF, ATF-IDF and Formal Concept Analysis algorithms was proposed by Hsieh, et al., [8]. The new approach first constructed as Concept Lattice with keywords extracted by ATF-IDF from documents to ensure a relationship hierarchy between keywords represented concepts. Then FCA was used to compute intra-document relationships to decide on a correct learning path.

Data classification for cross domains were researched and is a basic method to distinguish one from another, as it needs to know what belongs to which group. It can infer unseen dataset with unknown class through structural similarity analysis of a dataset with known classes. Classification results reliability is crucial. The higher the generated classification results accuracy, the better the classifier. They regularly seek to improve classification accuracy through either existing techniques or through developing new ones. Various procedures are used to improve classification accuracy performance. While most methods try to improve classifier techniques accuracy, Omar, et al., [9] reduced dataset features number by choosing only relevant features prior to handing over dataset to classifier. Thereby motivating need for methods capable of selecting relevant features with lowered information loss. The aim is to reduce classifier workload using feature selection. The review reveals that classification with feature selection produced impressive results with accuracy.

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

Feature selection has gained importance due to its contribution to save classification cost with regard to time/computation load. Searching for essential features, a feature search method is through decision trees. The latter is an intermediate feature space inducer to select essential features. Some studies used decision tree as feature ranker with direct threshold measure in decision tree based features selection, while others remain decision trees but use pruning which acts as a threshold mechanism in feature selection. Yacob, et al., [10] suggested a threshold measure using Manhattan Hierarchical Cluster distance for use in feature ranking to select relevant features as part of feature selection procedure. Results were promising and can be further improved by adding higher number of attributes test cases.

Feature selection reduces features number in applications where data has 100's/1000's of features. Present feature selection focuses on locating relevant features. Yu and Liu [11] demonstrated that feature relevance is insufficient to ensure efficient high dimensional data feature selection. Feature redundancy was defined/proposed to perform feature selection redundancy analysis. A new framework decoupling relevance analysis and redundancy analysis was proposed. A correlation-based method for relevance/redundancy analysis was developed and studied its efficiency/effectiveness compared to representative methods.

Principal component analysis (PCA) is the mainstay of data analysis - a black box used and usually poorly understood. Shlens [12] dispelled this myth as the manuscript aimed to build a solid intuition for how/why PCA works. It crystallized this knowledge by deriving the mathematics behind PCA from simple intuitions It was felt that by addressing all aspects, all readers would have an improved PCA understanding and also the when, how and why of this technique's application.

A new matrix learning scheme extending the Relevance Learning Vector Quantization (RLVQ), to a general adaptive metric was proposed by Schneider, et al., [13]. By introducing a full relevance factors matrix in distance measure, correlations between features and classification scheme importance are considered and automated, a general metric adaptation happens during training. Compared to weighted Euclidean metric used in RLVQ and its variations, a total matrix powerfully represents data's internal structure correctly. Large margin generalization bounds are transferred to this, leading to input dimensionality independent bounds. This includes local metrics attached to all prototypes corresponding to piecewise quadratic decision boundaries. The algorithm was tested and compared to alternative LVQ schemes using artificial data set, benchmark UCI repository multi-class, and an issue from bioinformatics, recognition of splice sites for C.

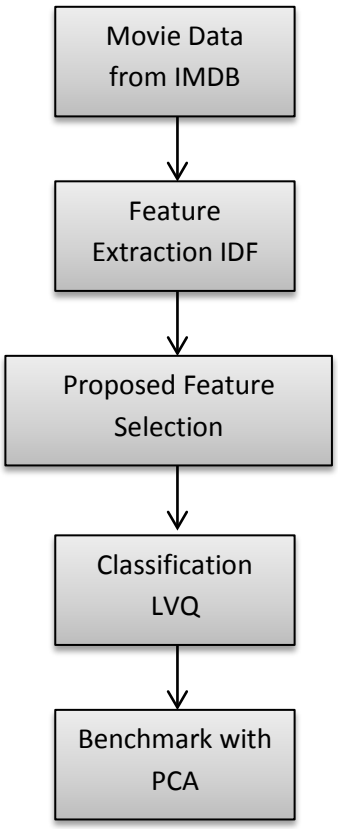

Figure 1: Flowchart Of Proposed Method

3. METHODOLOGY

The flowchart of the proposed methodology is shown in Figure 1 and the following sections details the steps in the proposed methodology.

3.1 IMDb Database

The IMDb is a large database with relevant and comprehensive information on movies- past, present and future [14]. It began as a shell scripts set and data files. The latter was a collection of email messages between users of rec.arts.movies Usenet bulletin board. Such movie fans exchanged information on actors,

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

actresses and directors and also biographical information on moviemakers. At some point, such data files became searchable with commands built by shell scripts.

IMDb uses two methods to add information to a database: Web forms and e-mail forms. Information from submission procedures indicates that, it is simpler to use web forms rather than e-mail format, if only addition to information is an update. If new information is to be submitted, users request or obtain format templates from IMDb through e-mail. The proposed information has to be formatted according to templates and validated.

3.2 Inverse Document Frequency (IDF)

Inverse document frequency (IDF) is a numerical statistic showing the importance of a word to a document, in a collection/corpus [15]. It is used as a weighting factor in information retrieval/text mining. IDF value increases with the repeated appearance of a word in a document. But offset by word's frequency in the corpus, which controls the fact that some words are more common than others. IDF weighting scheme variations are used by search engines as central tool to score and rank a document's relevance given a user query. IDF is used for stop-words filtering in subject fields including text summarization/classification. Text Classification is a semi-supervised machine learning task automatically assigning a document to a pre-defined categories set based on textual content, extracted features.

IDF appears in many heuristic measures of information retrieval. But till date IDF has been a heuristic itself. It is defined as a logarithm of the ratio of documents number containing a given word. Rare words have high IDF while common function words like "the" have low IDF. IDF measures a word's ability to discriminate documents. Text Classification assigns a text document to a pre-defined class set automatically, using machine learning. Classification is based on significant words/key-features of text document. As classes are pre-defined, it is a supervised machine learning process.

The term document frequency is computed as follows: for a document set x and a set of terms a. A document is modelled as a vector v in a dimensional space R^a. When term frequency denoted by $freq(x,a)$, expresses number of occurrences of term a in document x. Term-frequency matrix $TF(x,a)$ measures term association a regarding a given document x. $TF(x,a)$ is assigned zero when document has no term and $TF(x,a)$ = 1when term a occurs in document x or uses relative term frequency; term frequency as against total occurrences of document terms. Frequency is generally normalized by (Liu, et al., 2007):

$$TF(x,a)=\begin{cases}0 & freq(x,a)=0\\ 1+\log\left(1+\log\left(freq(x,a)\right)\right) & otherwise\end{cases}$$

Inverse Document Frequency (IDF) represents scaling. When a term a occurs frequently in documents, its importance is scaled down due to lowered discriminative power. The $IDF(a)$ is defined as follows:

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

x_a is documents set having term a.

Though TF-IDF is a common metric in text categorisation, its use in sentiment analysis is not known much. It has been used as a unigram feature weight. TF-IDF has 2 scores, term frequency and inverse document frequency. Term frequency counts the many times a term occurs in a document, whereas inverse document-frequency is the result of dividing total documents by documents where a specific word appears repeatedly. Multiplication of these values leads to high score for words appearing repeatedly in limited documents. Terms appearing frequently in all documents have a low score [21].

3.3 Proposed Feature Selection Based on Decision Trees

Decision trees are popular methods for inductive inference. They are robust to noisy data and learn disjunctive expressions. A decision tree is a k-array tree in which each internal node specifies a test on some attributes from input feature set representing data. Each branch from a node corresponds to possible feature values specified at that node. And every test results in branches, representing varied test outcomes. The

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

decision tree induction basic algorithm is a greedy algorithm constructing decision trees in a top-down recursive divide-and-conquer manner [16].

The algorithm begins with tuples in the training set, selecting best attribute yielding maximum information for classification. It generates a test node for this and then a top down decision trees induction divides current tuples set according to current test attribute values [17]. Classifier generation stops when all subset tuples belong to the same class or if it is not worthy to proceed with additional separation to further subsets, i.e. if more attribute tests yield information for classification alone below a pre-specified threshold. In this paper, it is proposed to base the threshold measure based on information gain and Manhattan hierarchical cluster.

In the proposed feature selection, a Decision tree induction selects relevant features. Decision tree induction is the learning of decision tree classifiers constructing tree structure where each internal node (no leaf node) denotes attribute test. Each branch represents test outcome and each external node (leaf node) denotes class prediction. At every node, the algorithm selects best partition data attribute to individual classes. The best attribute to partitioning is selected by attribute selection with Information gain. Attribute with highest information gain splits the attribute. Information gain of the attribute is found by

$$\inf o(D) = -\sum_{i=1}^{m} p_i \log_2 (p)$$

Where p_i is the probability that arbitrary vector in D belongs to class c_i. A log function to base 2 is used, as information is encoded in bits. Info (D) is just average information amount required to identify vector D class label. The information gain is used to rank the features and the ranked features are treated as features in hierarchical clusters. The proposed Manhattan distance for n number of clusters is given as follows:

$$MDist = \sum_{i=1}^{n} (a_i - b_i)$$

A cubic polynomial equation is derived using the Manhattan values and the threshold criterion is determined from the slope of the polynomial equation. The features are assumed to be irrelevant for classifying if the slope is zero or negative and relevant when the slope is positive.

3.4 Principal Component Analysis

When input dimensions are large and components highly correlated, dimensions are reduced using PCA [18]. For a variable set, PCA calculates artificial variables smaller set representing observed variable's variance. Artificial variables calculated are principal components used as predictor, criterion variable in the analysis. PCA orthogonalises variables and resulting principal components with large variation and eliminates components with least variation from datasets. When applied on a dataset PCA observes the following steps.

1. Mean subtracted from each data dimensions producing a data set with zero mean.
2. Covariance matrix is calculated.
3. Eigenvectors and eigenvalues of the covariance matrix are calculated.
4. Highest eigenvalues are principal components of dataset. Remove eigenvalues of less significance to form feature vector.
5. A new dataset is derived.

3.5 Learning Vector Quantization (LVQ)

Learning Vector Quantization (LVQ) is a local classification algorithm, where classification boundaries are locally approximated, the difference being that instead of using all training dataset points, LVQ uses only a prototype vectors set. This ensures efficient classification as vectors number needing storing or comparing is reduced greatly. Additionally, a carefully chosen prototype set also increase noise problems in the classification accuracy [18].

LVQ is an algorithm that learns appropriate prototype positions used classification and is defined by P prototypes set $\{(m_j, c_j), j = 1\ldots P\}$, where m_j is a K-dimensional vector in feature space, and c_j its class label. The prototypes number is larger than classes number. Thus, each class is represented by more than one prototype. Given an unlabeled data point x_u, its class label y_u is determined as class c_q of nearest prototype m_q

$$y_u = c_q, q = \arg\min_j d(x_u, m_j)$$

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

Where *d* is Euclidean distance. Other distance measures are used depending on the problem.

4. RESULTS AND DISCUSSION

Features are extracted using IDF from the movie data. The PCA and the proposed feature selection method were used to reduce the features. Table 1 and Figure 2 shows the classification accuracy obtained from LVQ and compared with Naïve Bayes classifier and Classification and Regression Tree (CART). Table 2 and Figure 3 gives the Root Mean Squared Error (RMSE).

Table 1: Classification Accuracy

Technique used	Classification accuracy
A. CART with PCA	57.51
B. Naïve Bayes with PCA	70.01
C. Naïve Bayes with LVQ	75.02
D. CART with proposed feature extraction	65.75
E. Naïve Bayes with proposed feature extraction	75.51
F. Naïve Bayes with LVQ and proposed feature extraction	**79.75**

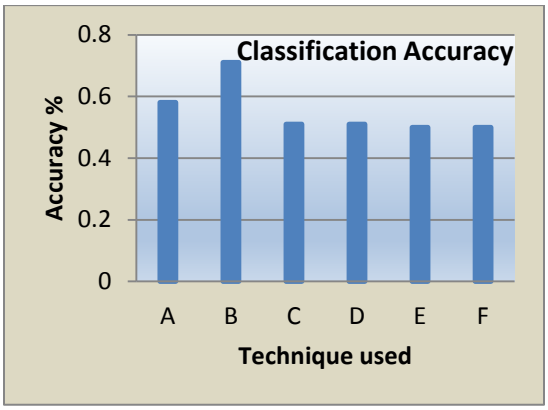

Figure 2 Classification Accuracy

It can be seen from figure 2, the classification accuracy obtained through Naïve Bayes with LVQ is better than Naïve Bayes with PCA by around 5%. Figure 3 shows the Root Mean Squared Error (RMSE).

Table 2: Root Mean Squared Error

Technique used	RMSE
A. CART with PCA	0.61
B. Naïve Bayes with PCA	0.54
C. Naïve Bayes with LVQ	0.54
D. CART with proposed feature extraction	0.44
E. Naïve Bayes with proposed feature extraction	0.41
F. Naïve Bayes with LVQ and proposed feature extraction	0.36

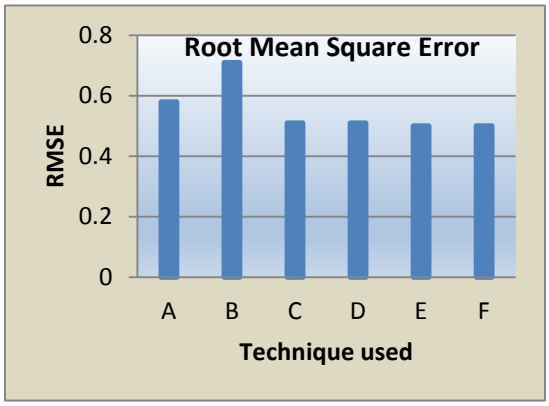

Figure 3: Root Mean Squared Error

Table 3: Precision And Recall

Technique used	Precision	Recall
A. CART with PCA	0.58	0.538
B. Naïve Bayes with PCA	0.71	0.714
C. Naïve Bayes with LVQ	0.51	0.773
D. CART with proposed feature extraction	0.51	0.669
E. Naïve Bayes with proposed feature extraction	0.50	0.715
F. Naïve Bayes with LVQ and proposed feature extraction	0.50	0.799

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

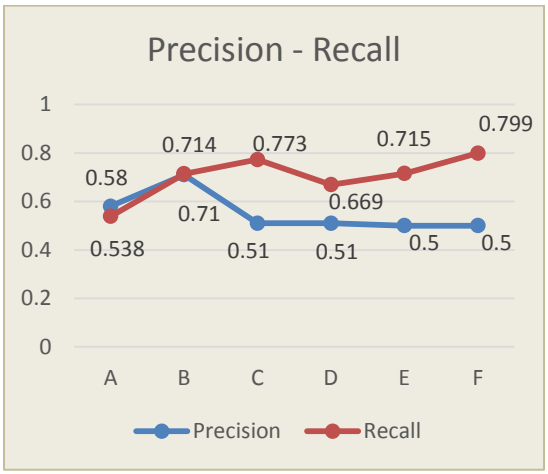

Figure 4: Precision & Recall

It can be seen that the precision and recall low for the three classifiers.

5. CONCLUSION

Rapid advances in computer based high-throughput technique provided unparalleled chances for humans to expand production, services, communications, and research productions. Meanwhile, immense high-dimensional data quantities accumulate challenging state-of-the-art data mining techniques. Feature selection is needed for successful data mining applications, as they lower data dimensionality removing irrelevant features. In this paper, a feature selection for Opinion mining using decision tree is proposed. LVQ type learning models constitute popular learning algorithms due to their simple learning rule, their intuitive formulation of a classifier by means of prototypical locations in the data space, and their efficient applicability to any given number of classes. Movie review features obtained from IMDb was extracted using inverse document frequency and the importance of the word found. Principal component analysis was used for feature selection based on the importance of the work with respect to the entire document. The classification accuracy obtained by LVQ was 75%. However it was observed that the precision for positive opinions was quite low. This phenomenon was observed not only on LVQ but other classifiers including CART and Naïve Bayes.

REFERENCES:

[1]. 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.

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

[3]. Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, *2*, 568.

[4]. Somprasertsri, G., & Lalitrojwong, P. (2010). Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization. *J. UCS*, *16*(6), 938-955.

[5]. 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.

[6]. Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., & Liu, H. (2010). Advancing feature selection research. *ASU Feature Selection Repository*.

[7]. El-Halees, A., & Gaza, P. (2011). Mining Feature-opinion in Educational Data for Course Improvement. *International Journal of New Computer Architectures and their Applications (IJNCAA)*, *1*(4), 1076-1085.

[8]. Hsieh, T. C., Chiu, T. K., & Wang, T. I. (2008, July). An approach for constructing suitable learning path for documents occasionally collected from internet. In *Machine Learning and Cybernetics, 2008 International Conference on* (Vol. 6, pp. 3138-3143). IEEE.

[9]. Omar, N., Jusoh, F., Ibrahim, R., & Othman, M. S. (2013). Review of Feature Selection for Solving Classification Problems. *JISRI*, *3*.

[10]. Yacob, Y. M., Sakim, H. M., & Isa, N. M. (2012). Decision tree-based feature ranking using manhattan hierarchical cluster criterion. *International Journal of Engineering and Physical Sciences*, *6*.

[11]. Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and

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

redundancy. *The Journal of Machine Learning Research*, *5*, 1205-1224.

[12]. Shlens, J. (2005). A tutorial on principal component analysis. *Systems Neurobiology Laboratory, University of California at San Diego*.

[13]. Schneider, P., Biehl, M., & Hammer, B. (2009). Adaptive relevance matrices in learning vector quantization. *Neural Computation*, *21*(12), 3532-3561.

[14]. The Internet Movie Database Ltd. Internet movie database. http://www.imdb.com.

[15]. Metzler, D. (2008, October). Generalized inverse document frequency. In*Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 399-408)

[16]. Ratanamahatana, C. A., & Gunopulos, D. (2002). Scaling up the naive bayesian classifier: Using decision trees for feature selection.

[17]. Gayatri, N., Nickolas, S., & Reddy, A. V. (2010). Feature selection using decision tree induction in class level metrics dataset for software defect predictions. In *Proceedings of the World Congress on Engineering and Computer Science* (Vol. 1, pp. 124-129).

[18]. Friston, K. J., Frith, C. D., Liddle, P. F., & Frackowiak, R. S. J. (1993). Functional connectivity: the principal-component analysis of large (PET) data sets. Journal of cerebral blood flow and metabolism, 13, 5-5.

[19]. Grbovic, M., & Vucetic, S. (2009, June). Learning vector quantization with adaptive prototype addition and removal. In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on* (pp. 994-1001). IEEE.

[20]. Jeevanandam Jotheeswaran et al., (2012), Feature Reduction using Principal Component Analysis for Opinion Mining. IJCST, Volume 3, Issue 5, May 2012 (P 118 – 121).