Research Journal of Applied Sciences, Engineering and Technology 9(7): 519-525, 2015
DOI:10.19026/rjaset.9.1435
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2015 Maxwell Scientific Publication Corp.
Submitted: September 29, 2014 Accepted: November 3, 2014 Published: March 05, 2015

Research Article

Framework for Evaluating Camera Opinions

[1]K.M. Subramanian and [2]K. Venkatachalam
[1]Department of Computer Science Engineering, Erode Sengunthar Engineering College,
[2]Department of Electronics and Communication Engineering, Velalar College of Engineering and Technology, Erode, Tamilnadu, India

Abstract: Opinion mining plays a most important role in text mining applications in brand and product positioning, customer relationship management, consumer attitude detection and market research. The applications lead to new generation of companies/products meant for online market perception, online content monitoring and reputation management. Expansion of the web inspires users to contribute/express opinions via blogs, videos and social networking sites. Such platforms provide valuable information for analysis of sentiment pertaining a product or service. This study investigates the performance of various feature extraction methods and classification algorithm for opinion mining. Opinions expressed in Amazon website for cameras are collected and used for evaluation. Features are extracted from the opinions using Term Document Frequency and Inverse Document Frequency (TDF×IDF). Feature transformation is achieved through Principal Component Analysis (PCA) and kernel PCA. Naïve Bayes, K Nearest Neighbor and Classification and Regression Trees (CART) classification algorithms classify the features extracted.

Keywords: K nearest neighbor and Classification and Regression Trees (CART), naïve bayes, opinion mining, Principal Component Analysis (PCA) and kernel PCA, TDF×IDF

INTRODUCTION

Opinion mining in textual materials like Weblogs is another technologies dimension facilitating search and summarization. Opinion mining identifies author's viewpoint on a subject instead of just identifying subject alone. Present approaches divide problem space into sub-problems. For example, creating a useful features lexicon classifies sentences into positive, negative or neutral categories. Present techniques identify words, phrases and patterns indicating viewpoints (Conrad and Schilder, 2007). This was difficult, as it is not just a keyword which matters, but the context. For example, this is a great decision, reveals clear sentiment and but that the decision announcement produced much media attention is neutral.

Opinion mining is also termed as sentiment analysis/sentiment classification. Opinion mining emphasis is not on topic of the text, but the author's attitude to the topic. Recently, opinion mining was applied to movie reviews, commercial products and services reviews, to Weblogs and to News. Such subtasks include.

Subjectivity analysis: Involves determining if a text is objective or subjective; this is also a binary classification task.

Polarity analysis: Includes predicting whether a text established as subjective is positive or negative in polarity.

Polarity degree: Measures polarity degree, positive/negative in subjective text.

Generally, opinions are expressed on anything, e.g., a product, service, topic, individual, organization, or event. The term object denotes the entity commented on. An object has components (or parts) and attributes. Each component also has sub-components and attributes. Thus, based on part-of relationship an object can be hierarchically decomposed.

Definition (object): An object O is a unit which is a product, event, person, organization or topic. It is connected with a pair, O: (T, A), where T is components (or parts) hierarchy or taxonomy and O's sub-components and A an attributes set of O. Each component has own sub-components and attributes sets.

Definition (opinion passage on a feature): A feature f opinion passage of object O evaluated in d is a

Corresponding Author: K.M. Subramanian, Department of Computer Science Engineering, Erode Sengunthar Engineering College, Erode, Tamilnadu, India

consecutive sentences group in d expressing positive/negative opinion on f. It is possible that a single sentence states opinions on more than one feature, e.g., “This camera’s picture quality is good, but has a short battery life”.

Definition (opinion holder): The holder of a specific opinion is a person/organization holding that opinion. In product reviews, forum postings and blogs, opinion holders are authors of posts (Hu and Liu, 2004).

Online reviews express opinions about a product or service and users evaluate a product or service based on these opinions before buying or using the product. Due to the huge amount of reviews available in different websites, it is hard to comprehend all the opinions. Opinion mining summarizes and the polarity of the various reviews which helps in gaining a overall picture about a product or service. The Sentiment is classified as negative, neutral or positive on retrieving the information from the review. Various techniques such as clustering, supervised learning methods classify sentiment polarity (Liu and Zhang, 2012). Sentiment classification has been widely researched and several approaches are surveyed in literature (Baccianella *et al*., 2010; Cambria *et al*., 2013).

This study investigates the efficacy of the feature extraction methods and classification algorithms for classifying cameras reviews. Opinions expressed on cameras are taken from Amazon website. TDF×IDF is used for extracting features from the camera reviews. Feature transformation is undertaken by using PCA and kernel PCA. Naïve Bayes and K Nearest neighbour classifiers and CART algorithms performance evaluation s investigated.

LITERATURE REVIEW

Samsudin *et al*. (2011) proposed Bess or xbest mining Malaysian online reviews where opinion mining of online movie reviews from many for and blogs written by Malaysians is studied. Experiment data was tested using machine learning classifiers like Support Vector Machine (SVM), Naïve Bayes and k-Nearest Neighbor (kNN). The result illustrated that machine learning techniques performance without preprocessing of micro-texts/feature selection was low. Hence, additional steps were required to mine opinions from data.

Research on Internet Public Opinion analysis technology based on topic cluster was proposed by Chunhua *et al*. (2010) where Internet users search data cluster with K-nearest neighbor and shortest path approaches undertaken. It formed an association search net and provided shortest path. It analyzed Internet user's search behavior and characteristics. Finally it discovered information dissemination pattern guiding Internet public opinion trends correctly.

Internet public opinion research tracking algorithm was proposed by Lu and Yao (2011) describing information means about internet public opinion and study situation. It analyzed internet public opinion's tracking algorithm SVM, KNN and NB.

A sequential feature extraction approach to Naive Bayes classification of microarray data was proposed by Fan *et al*. (2009) consisting of feature selection through stepwise regression and feature transformation through class conditional independent component analysis. Experiment results on five microarray datasets proved the proposed approach’s effectiveness in

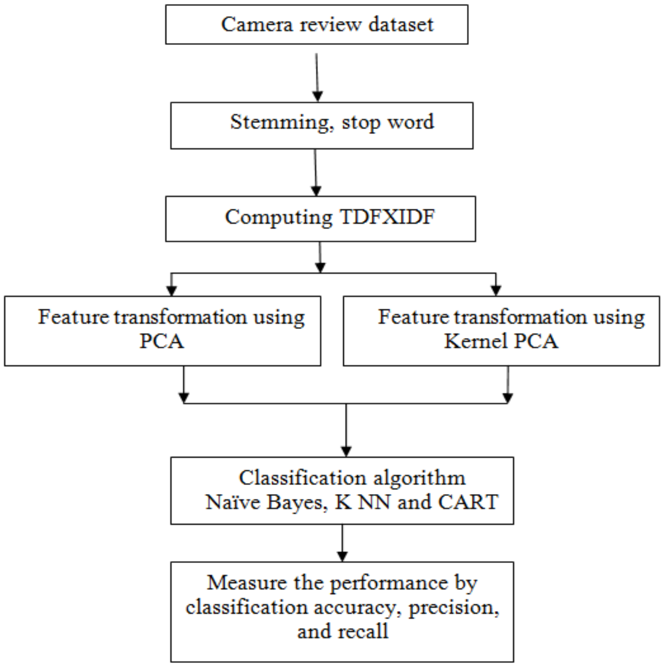
Camera review dataset
Stemming, stop word
Computing TDFXIDF
Feature transformation using PCA
Feature transformation using Kernel PCA
Classification algorithm Naïve Bayes, K NN and CART
Measure the performance by classification accuracy, precision, and recall

Fig. 1: Flowchart of the methodology

improving performance of naive Bayes classifier in microarray data analysis.

Opinion Mining Classification Using Key Word Summarization based on Singular Value decomposition was suggested by Valarmathi and Palanisamy (2011). This method aimed to develop a method using Singular Value Decomposition based word score by modeling a custom corpus for a topic where opinion mining is planned. Bayes Net and decision tree induction algorithms classified opinions.

METHODOLOGY

This study investigates opinion mining for camera reviews. TDF×IDF is used for feature extraction. Feature transformation is by using PCA and kernel PCA. Naïve Bayes, K Nearest neighbor and CART algorithms study accuracy.

The flowchart of the methodology followed is shown in Fig. 1.

Camera dataset: Opinions are collected from amazon http://personalwebs.coloradocollege.edu/~mwhitehead/htmL/opinion_mining.htmL. Two hundred and twenty five each of positive and negative reviews are used. Some examples of the positive and negative reviews are presented here.

Positive reviews: ‘We bought this camera and have been more than happy with it's performance. We are not professional photographers; we just needed something easy, cheap and reliable. This camera is all those things! The battery issue we heard about does not seem to be a problem, the pictures come out crisp and it could not be easier to learn how to use. We are very pleased with this product.’

‘This is my second Sony cybershot digital camera, although I have purchased Kodak Easy Shares for family members. I loved the first one (only 3.2 MGP). This camera is perfect for the not-so-tech-wise consumer. It takes great pictures and has a high quality Zeiss lens. Most of all it is easy to use, especially for the beginner and intermediate user. It stores easily in a pocket and I love the color choices Sony gives you! The review pictures button is a little small, but you get used to it quickly.’

This is a fantastic little camera-especially for point and shoot users who just want a camera to take snapshots and is not interested in becoming a rocket scientist in order to learn how to operate the camera. A 7.2 MP and fast shutter with reasonable flash for its size, its hard to mess up pictures.’

Negative reviews: ‘Battery life is terrible if you use image stabilizer, expect 25-30 shots on a full battery. Also the camera lacks of an optical view finder very difficult to shoot in sunlight with LCD. Many shots are somewhat bleached out while using auto white balance. Owned a stylus 400 digital prior to this and it is a disappointment rather than upgrade. The only upsides are the 5x optical zoom which is a little choppy and the image stabilizer that kills the battery if left on.’

‘Bulkier than it looks and it feels like a toy. Not very solid at all and the pics aren't that amazing either.’

‘I purchased this camera to snap off some photos when I moved to Vancouver for school (I left my other camera on the other side of the country) and it's one of the worst mistakes I've ever made. When it's not taking blown out white pictures or pitch black images, it's snapping off blurry or orange tinted images. I brought my first one back for another one-the same problem! (And before anyone says that I just don't know how to use it, keep in mind that I've been a photographer for a few years.) Add to that the countless number of reviews for this camera for the same problems that I'm having and you get one bottom line-THIS CAMERA IS A DUD! I'll never buy another Sony camera again in my life. In fact, as the go-to-guy for my friends when buying tech gear, I've told them to stay away from Sony cameras from introductory to professional. I'm sticking with my Nikon from now on. This is possibly the worst camera I've ever used (and that says a lot)’

Stemming: Stemming is a reference to root word origins. For example, search is the root term for Search, Searching and Searches. In many cases, words morphological variants have similar semantic interpretations and are considered equal for IR applications. Due to this reason, many so called stemming Algorithms, or stemmers, were developed to reduce a word to its stem or root form. Thus a query or document’s key terms are represented by stems and not by original words. This means that a term’s differing variants can be conflated to single representative form-in addition to reducing dictionary size, that is, number of distinct terms required to represent a documents (Das and Bandyopadhyay, 2010) set.

Stop word: A general stop word list for words without purpose for retrieval, but frequently used to compose documents, are developed for two main reasons: First, it is possible that a query and document match is based on good indexing terms. So, retrieving document which has words like "be", "the" and "your" in corresponding request is not intelligent strategy. These non-significant words represent noise and damage retrieval performance failing to discriminate between relevant and non-relevant documents. Secondly, it is expected to reduce inverted file size to a range between 30 and 50% (Savoy, 1999).

The occurrences of every word in a document are represented through Term Frequency (TF) that is a document specific measure of term importance. A documents collection being considered is a corpus. Many term weighting techniques were proposed in the literature. A document vector represents a vector space model whose components are term weights. A document using term frequency as term weights is represented in vector form as $\{tf_1, tf_2, tf_3, \dots, tf_n\}$, where tf is term frequency and n total terms number in document.

Document length in a corpus varies with longer documents having higher term frequencies and unique terms compared to shorter documents. Cosine function is measures similarity between two documents. It is given by:

$$cos\left(d_i, d_j\right) = \frac{d_i.d_j}{\|d_i\|X\|d_j\|} \tag{1}$$

where, d_i denotes the i^{th}document vector.

As term frequency favors long documents because of higher term frequencies, it is suggested to normalize term frequency of jth term through maximum term frequency in same document:

$$TF_j = \frac{(tf\,j)}{(size(d))} \tag{2}$$

While TF is a term's importance local measure, Inverse Document Frequency (IDF) is global used to show corpus term importance. It assigns lesser values to words in most documents and higher values to those in fewer documents (Jotheeswaran *et al*., 2012).

When dataset documents are modelled as vector v, for a set of documents x and a terms a, in dimensional $space^R$ it is a vector space model. When a term 'a' occurs in document x, number of occurrences of term is given through term frequency denoted by $freq(x, a)$ the term association regarding a given document x is measured by term-frequency matrix TF (x, a). Term frequencies are given values based term occurrence, so TF (x, a) is assigned either zero if document does not have term or a number. The number can be set as TF (x, a) = 1 when term 'a' is in document x or uses relative term frequency. Relative term frequency is term frequency versus total occurrences of all terms in a document. Term frequency is normalized by equation:

$$TF(x, a) = \begin{cases} 0\, freq(x, a) = 0 \\ 1 + \log\left(1 + log\left(freq(x, a)\right)\right) \end{cases} \tag{3}$$

Inverse Document Frequency (IDF) represents scaling. Importance of term 'a' is scaled down if term occurs frequently in documents due to lowered discriminative power (Isabella and Suresh, 2012). IDF (a) is defined as equation:

$$IDF(a) = \log\frac{1+|x|}{x_a} \tag{4}$$

x_a = The set of documents containing term a.

Combining term frequency and inverse document frequency is called TFIDF used to represent term weight numerically:

$$TFIDF\ =\ TF \times IDF \tag{5}$$

The weight for a term i as regards of TF-IDF is given by:

$$W_i = \frac{\left(TF_iX\log\left(\frac{N}{n_i}\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(TF_iX\ \log\left(\frac{N}{n_i}\right)^2\right)}} \tag{6}$$

where,

N = Number of total documents
n_i = Document frequency of term i

If a document contains 120 words where the word lens appears 4 times, then TF for lens is (4/120) = 0.033. If the total dataset has 10 million documents and the word lens appears in one thousand of these. Then, the IDF is calculated as log (10,000,000/1,000) = 4. Thus, the TF-IDF weight is the product of these quantities: 0.033 * 4 = 0.132.

Principal Component Analysis (PCA): Principal Component Analysis (PCA) is a technique to dimensionally reduce and extract features. PCA tries to find lower dimensionality linear subspace of original feature space where new features have largest variance This is called dimensionality reduction, as vector $\bar{x}$ containing original data and is N-dimensional is lowered to a compressed vector $\bar{c}$ that is M-dimensional, where M<N. A vector $\bar{x}$ is coded into a vector $\bar{c}$ with reduced dimension. Vector $\bar{c}$ is stored, transmitted or processed resulting in vector $\bar{c}'$, capable of being decoded back to a vector $\tilde{\bar{x}}'$. The last vector is a result approximation which can be reached by storing, transmitting or processing vector $\bar{x}$ (Jolliffe, 2005) (Fig. 2).

The diagram's encoder should perform a linear operation, using a matrix $\bar{\bar{Q}}$:

$$\bar{c} = \overline{\overline{Q}}\,\overline{x} \tag{7}$$

Decoder is also a linear operation, written as a sum of vector elements of $\bar{c}$ multiplied by matrix columns:

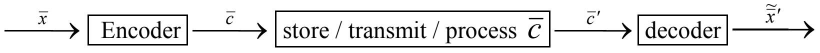

Fig. 2: Process of PCA

$$\overline{\overline{Q}} \ : \ \overline{\widetilde{x}} = \overline{c}^T \overline{\overline{Q}}^T \rightarrow \overline{\widetilde{x}} = \sum_{i=1}^{M} c_i \overline{q}_i \tag{8}$$

Kernel PCA: Traditional PCA permits linear dimensionality reduction. But, if data includes complicated structures that cannot be simplified in linear subspace, traditional PCA becomes invalid. But, kernel PCA permits generalization of traditional PCA to nonlinear dimensional reduction.

Kernel Principal Component Analysis (kernel PCA) as a nonlinear generalization of principal component analysis was introduced in Honkela *et al.* (2004) the aim being to map given data points from input space $\mathbb{R}^n$ to high-dimensional (infinite-dimensional) feature space $\mathcal{F}$:

$$\Phi = \mathbb{R}^n \rightarrow \mathcal{F} \tag{9}$$

and perform PCA in F. The space F and also mapping Φ might be complicated. But using so-called kernel trick, it avoids using Φ explicitly: PCA in F is formulated so that only F's inner product is needed which is seen as a nonlinear function called kernel function:

$$\mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R} \tag{10}$$

$$(x, y) \rightarrow k(x, y) \tag{11}$$

This calculates each pair of vector's real number from input space.

Naive bayes classifier: Naïve Bayes are statistical classifier based on Bayes theorem (McCallum and Nigam, 1998) which uses a probabilistic approach to predict given data's class matching it to the class with highest posterior probability. Following are Naïve Bayes algorithms:

$$P(C_i|V) = \frac{P(V|C_i)P(C_i)}{P(V)} \tag{12}$$

where, $V = (v_1, \ldots\ldots, v_n)$ is document represented in n-dimensional attribute vector and $c_1, \ldots\ldots, c_m$ represents m class. But it is computationally expensive to compute $P\ (V|C_i)$. To reduce computation, naïve conditional independence assumption of class is made. Thus:

$$P(V|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{13}$$

K-nearest neighbour classification: *k*-Nearest neighbour classifier is based on premises that vector space model is similar for similar documents. Training documents are indexed and each associated with corresponding label. A submitted test document is treated like a query retrieving from training set, documents similar to test document. The test document class label is assigned based on distribution of k nearest neighbours. Class label can be refined by adding weights. Tuning k, obtains higher accuracy. Nearest neighbour method is easy to understand and implement (Kulkarni *et al.*, 1998):

$$p(x) \cong \frac{k}{NV} \tag{14}$$

Similarly, probability density function p (x|Hi) of observation x conditioned to hypothesis H_i is approximated 24. Let us assume $N_{i\_}$ is number of patterns associated to hypothesis:

$$H_i, i = 1 \ldots C, so\ that\ N1 + \cdot \cdot \cdot \cdot + NC = N \tag{15}$$

Classification and Regression Trees (CART): Classification and Regression Trees (CART) handles numerical and categorical variables. Among CART's advantages is its robustness to outliers. Usually splitting algorithm isolates outliers in individual node/nodes. A CART practical property is that classification or regression trees structure is invariant regarding independent variables monotone transformations. Any variable can be replaced with its logarithm or square root value and tree structure does not change (Timofeev, 2004):

$$i(t) - p_L i(t_L) - p_R i(t_R) \tag{16}$$

CART selects split maximizing impurity decrease
CART methodology has three parts:

- Maximum tree construction
- Choice of correct tree size
- New data classification using constructed tree

RESULTS AND DISCUSSION

The opinions are collected from Amazon website and 225 positive and 225 negative features are used in this study. Features are extracted using TDF×IDF and Feature transformation is achieved using PCA and kernel PCA. Accuracy of Naïve Bayes, K Nearest neighbour and CART algorithms to classify the reviews is evaluated. Experiments are conducted for:

- Feature extraction using only TDF×IDF
- Feature extraction using TDF×IDF and PCA
- Feature extraction using TDF×IDF and kernel PCA

Results obtained for classification accuracy are listed in Table 1.

Table 1: Classification accuracy for various methods

	TDF×IDF	TDF×IDF and PCA	TDF×IDF and kernel PCA
Naïve bayes	0.7422	0.7556	0.7627
CART	0.7733	0.7844	0.7911
KNN	0.7378	0.7467	0.7578

Table 2: Precision

	TDF×IDF	TDF×IDF and PCA	TDF×IDF and kernel PCA
Naïve bayes	0.74295	0.75555	0.76275
CART	0.74295	0.78490	0.79195
KNN	0.73785	0.74670	0.75775

Table 3: Recall

	TDF×IDF	TDF×IDF and PCA	TDF×IDF and kernel PCA
Naïve bayes	0.74225	0.75555	0.76275
CART	0.77335	0.78440	0.79110
KNN	0.73780	0.74665	0.75780

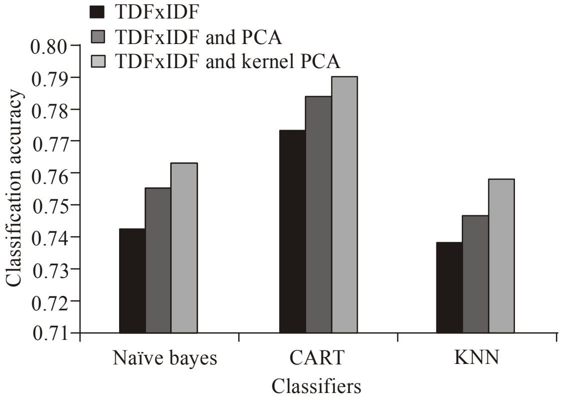

Fig. 3: Classification accuracy obtained for various methods

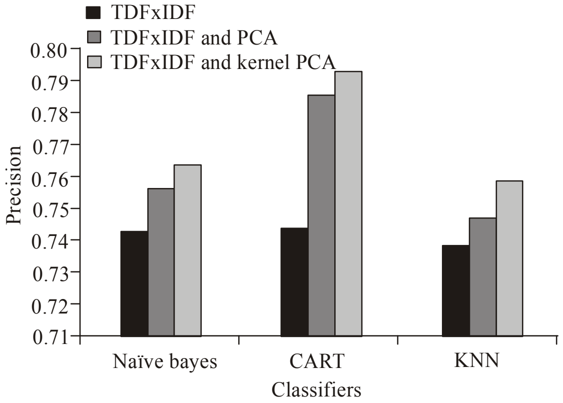

Fig. 4: Average precision obtained for various methods

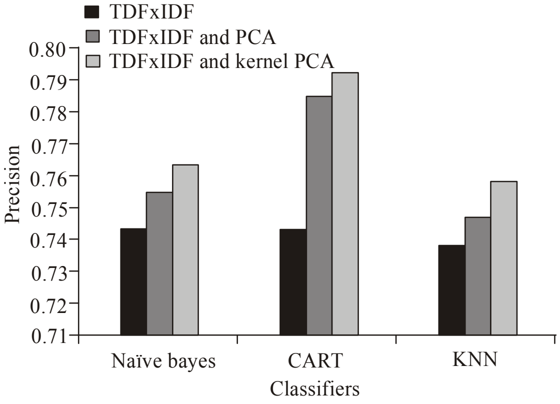

Fig. 5: Average recall obtained for various methods

It is observed from Table 1 and Fig. 3 that the CART achieves the best accuracy 79.11% for features selected using TDF×IDF and kernel PCA which is better by 3.72% when compared to Naïve Bayes and 4.39% when compared to KNN.

Table 2 and 3 tabulates the precision and recall achieved by various methods. Figure 4 and 5 depicts the precision and recall respectively.

It is observed from Table 3 and Fig. 4 that the CART with features selected using TDF×IDF and kernel PCA achieves the best precision of 0.792 which is better by 3.83% when compared to Naïve Bayes and 4.51% when compared to KNN.

Similar to precision, recall for the CART with features selected using TDF×IDF and kernel PCA achieves the best result of 0.791 which is better by 3.72% when compared to Naïve Bayes and 4.39% when compared to KNN.

CONCLUSION

A big part of information-gathering behavior is to find what people think. With availability and popularity of opinion-rich resources like online review sites and personal blogs, more chances and challenges arise as people now can and do use information technologies to understand others opinions. This study investigates the efficacy of the feature extraction methods and classification algorithms for classifying camera reviews. Reviews on camera are obtained from Amazon website. Feature from the reviews are extracted using TDF×IDF. Features are transformed using PCA and kernel PCA. Naïve Bayes and K Nearest neighbour classifiers and CART algorithms classify the features as positive or negative. Experimental results demonstrate that features extracted using TDF×IDF with kernel PCA improves the classification accuracy of the classifiers. The results reveal that CART algorithm has higher classification accuracy than other classifiers.

REFERENCES

Baccianella, S., A. Esuli and F. Sebastiani, 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Proceeding of the 7th Conference on International Language Resources and Evaluation (LREC, 2010), pp: 2200-2204.

Cambria, E., B. Schuller, X. Yunqing and C. Havasi, 2013. New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst., 28(2): 15-21.

Chunhua, Y., W. Shengwu, S. Yifan and Z. Gang, 2010. Research on analysis technology of Internet Public Opinion based on topic cluster. Proceeding of the 2nd International Conference on Information Science and Engineering (ICISE, 2010), pp: 6002-6005.

Conrad, J.G. and F. Schilder, 2007. Opinion mining in legal blogs. Proceeding of the 11th International Conference on Artificial Intelligence and Law, pp: 231-236.

Das, A. and S. Bandyopadhyay, 2010. Phrase-level polarity identification for Bengali. Int. J. Comput. Linguist. Appl., 1(1-2): 169-182.

Fan, L., K.L. Poh and P. Zhou, 2009. A sequential feature extraction approach for naïve bayes classification of microarray data. Expert Syst. Appl., 36(6): 9919-9923.

Honkela, A., S. Harmeling, L. Lundqvist and H. Valpola, 2004. Using kernel PCA for initialisation of variational Bayesian nonlinear blind source separation method. In: Puntonet, C.G. and A. Prieto (Eds.), ICA, 2004. LNCS 3195, Springer-Verlag, Berlin, Heidelberg, pp: 790-797.

Hu, M. and B. Liu, 2004. Mining opinion features in customer reviews. Proceeding of the National Conference on Artificial Intelligence. AAAI Press, MIT Press, Menlo Park, Cambridge, CA, MA, London, pp: 755-760.

Isabella, J. and R. Suresh, 2012. Analysis and evaluation of feature selectors in opinion mining. Indian J. Comput. Sci. Eng., 3(6).

Jolliffe, I., 2005. Principal Component Analysis. John Wiley and Sons Ltd., New York.

Jotheeswaran, J., R. Loganathan and B. MadhuSudhanan, 2012. Feature reduction using principal component analysis for opinion mining. Int. J. Comput. Sci. Telecomm., 3(5): 118-121.

Kulkarni, S., G. Lugosi and S. Venkatesh, 1998. Learning pattern classification: A survey. IEEE T. Inform. Theory, 44(6).

Liu, B. and L. Zhang, 2012. A survey of opinion mining and sentiment analysis. Min. Text Data, 2012: 415-463.

Lu, S. and C. Yao, 2011. The research of internet public opinion's tracking algorithm. Proceeding of the International Conference on Electric Information and Control Engineering (ICEICE, 2011), pp: 5536-5538.

McCallum, A. and K. Nigam, 1998. A comparison of event models for naive bayes text classification. Proceeding of the AAAI-98 Workshop on Learning for Text Categorization, 752: 41-48.

Samsudin, N., M. Puteh and A.R. Hamdan, 2011. Bess or xbest: Mining the Malaysian online reviews. Proceeding of the 3rd Conference on Data Mining and Optimization (DMO, 2011), pp: 38-43.

Savoy, J., 1999. A stemming procedure and stopword list for general French corpora. J. Am. Soc. Inform. Sci., 50(10): 944-952.

Timofeev, R., 2004. Classification and regression trees (cart) theory and applications. M.A. Thesis, CASE, Humboldt University, Berlin.

Valarmathi, B. and V. Palanisamy, 2011. Opinion mining classification using key word summarization based on singular value decomposition. Int. J. Comput. Sci. Eng., 3(1).