Research Journal of Applied Sciences, Engineering and Technology 13(3): 202-208, 2016
DOI:10.19026/rjaset.13.2931
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: September 26, 2015 Accepted: October 30, 2015 Published: August 05, 2016

Research Article

Framework for Evaluating Feature Selection in Opinion Mining

[1]E.A. Neeba and [2]S. Koteeswaran
[1]Department of Information Technology, Rajagiri School of Engineering and Technology, Kochi,
[2]Department of Computer Science Engineering, Vel Tech Dr.RR & Dr.SR Technical University, India

Abstract: Opinion mining is important in text mining applications in brand and product positioning, consumer attitude detection, customer relationship management and market research. Applications result in new generation companies, products for reputation management, online market perception and online content monitoring. Web expansion encourages users to contribute or express opinions through blogs, videos and social networking sites which provide information for sentiment analysis regarding a product or service. This study investigates various feature extraction methods performance and opinion mining classification algorithm. Evaluation is through the use of opinions from amazon.com with product reviews. Features extraction is from opinions using Term Document Frequency and Inverse Document Frequency (TDF×IDF). Feature transformation is through Principal Component Analysis (PCA) and kernel PCA. Feature selection techniques like Information Gain (IG), Mutual Information (MI) and Fisher Score select features. Extracted features are classified by Naïve Bayes, k Nearest Neighbour and Classification and Regression Trees (CART) classification algorithms.

Keywords: Fisher Score (FS), Information Gain (IG), Mutual Information (MI), Naïve Bayes, opinion mining, Principal Component Analysis (PCA), TDF x IDFF

INTRODUCTION

Opinion Mining also called sentiment analysis due to the huge volume of opinion rich web resources like discussion forum, review sites and blogs which are available digitally. Opinion mining is a type of natural language processing, tracking the mood of people regarding a specific product/topic providing automatic extraction of opinions, emotions and sentiments in text in addition to tracking attitudes and feelings on the net. People express views in blog posts, comments, reviews and tweets on varied topics. Tracking products or brands and determining whether they are viewed positively/negatively is done through the web. Opinion mining has different tasks, e.g., opinion extraction, sentiment analysis, sentiment mining, affect analysis, subjectivity analysis, emotion analysis and review mining. But, they all come under the sentiment analysis or opinion mining umbrella. Sentiment classification, opinion summarization and feature based sentiment classification are some research fields which dominate sentiment analysis.

Opinion mining is used in many ways. In marketing, it tracks an ad campaign's success rate or a new product launch, determining products or services popularity and its versions also tell us about demographics which like/dislike specific features (Vinodhini and Chandrasekaran, 2012). A review about a digital camera might be broadly positive, but be specifically negative about its weight. The vendor gets a clear picture of public opinion than surveys or focus groups, when such information is identified systematically (Buche *et al*., 2013).

Opinion mining has research interest with growing avenues (social networks and e-commerce websites) for people to express sentiments on the net. A typical sentiment-analysis application involves 3 key subtasks, like holder detection, target extraction and sentiment classification. A simple and most extensively studied sentiment classification case is sentiment polarity classification, which is classification of labelling a sentiment-oriented document's polarity as positive/negative/neutral. Sentiment classification is undertaken at document, sentence, phase, or word level (Wu and Gu, 2014).

This study investigates efficacy of feature selection methods to classify product reviews. TDF x IDF is used to extract features from camera reviews. Feature selection is through Information Gain (IG), Fisher Score (FS) Mutual Information (MI) and Principal Component Analysis (PCA). Naïve Bayes algorithm classifies extracted features.

Corresponding Author: E.A. Neeba, Department of Information Technology, Rajagiri School of Engineering and Technology, Kochi, India

LITERATURE REVIEW

Techniques and approaches promising to enable opinion-oriented information-seeking systems directly were covered by Pang and Lee (2008). This study focuses on methods addressing challenges raised by sentiment-aware applications related to traditional fact-based analysis. This study includes a summarization of the evaluative text and broader issues as regards manipulation, privacy and economic impact that development of opinion-oriented information-access services lead to. Future work discusses resources and benchmark datasets, while an evaluation campaign is also provided.

Pak and Paroubek (2010) who focussed on Twitter, a micro-blogging platform, for sentiment analysis reveal how to collect a corpus for sentiment analysis and opinion mining automatically. It performs linguistic analysis of a corpus and explains the discovered phenomena. This study builds a sentiment classifier through the corpus, which determines a document's positive/negative/neutral sentiments. Evaluations reveal the new techniques to be efficient performing better than earlier methods.

Ding *et al*. (2008) proposed a holistic lexicon-based approach to solving problems by exploiting natural language expressions external evidences/linguistic conventions which ensured handling context dependent opinion words, which created difficulties for current algorithms. This study deals with special words/phrases and language constructs that impact opinions based on linguistic patterns. It effectively aggregates multiple conflicting opinion words in sentences. A system named Opinion Observer based on a new technique was implemented. Results used a benchmark product review data set and additional reviews reveal that the new technique was highly effective, significantly outperforming current methods.

A new method to forecast stock returns by mining opinion and sentiment from Web forum messages was proposed by Duan and Zeng (2013). Opinion about stock prices rise and drop was first extracted from forum user's messages. Unhealthy sentiment is recognized by pattern matching. A Bayesian model incorporating opinion/unhealthy sentiment is established to infer relation between stock returns and opinion and sentiment combination. Compared to experiments on China share/stock market and Guba Web forum was done and results show that the new method was effective.

Bollegala *et al*. (2013) proposed a method to overcome problems in cross-domain sentiment classification. The paper using labelled data for source domains and unlabeled data for source/target domains created a sentiment sensitive distributional thesaurus. Sentiment sensitivity is achieved by incorporating document level sentiment labels in context vectors as a basis to measure distributional words similarity. The thesaurus expanded feature vectors during training/testing in binary classifiers. The new method outperformed baselines and returned results comparable to earlier cross-domain sentiment classification methods on a benchmark data set with Amazon user reviews for products. Comparisons against a word polarity lexical resource SentiWordNet, shows that the created sentiment-sensitive thesaurus captures words accurately expressing similar sentiments.

Opinion tree, a new kind of tree defined flexible opinion mining models was proposed by Ding *et al*. (2009). Medium, coarse-grained and fine-grained (three different granularities) opinion mining are realized in a flexible, unified model. Flexible opinion mining procedure is for public opinions on the internet. Experiments showed that when opinion tree building was finished, overall opinion about an internet hot topic is formed in the opinion tree.

Chen *et al*. (2013) proposed a method to handle sentiment analysis for Cantonese opinion mining. Researchers put effects on mining user opinions from reviews. Research works on opinion mining though numerous presented methods which focus on handling opinion mining for English, Chinese, Japanese and other languages. There is nothing on how to conduct Cantonese Opinion Mining are lacking.

Cho *et al*. (2010) proposed a different opinion mining method from current ones. Present opinion mining techniques comprise feature-based opinion mining, sentiment classification, opinion spam, summarization, comparative sentence and relation mining, opinion search and linguistic dictionary construction like WordNet. The methods enable credibility evaluation and result in conversion influencing every opinion holder on the Internet and personal information, which are LIWC result analysis, including background information and tendency.

Liu *et al*. (2012) proposed a new method to deal with feature-level opinion mining problems which considers explicit/implicit features and opinion words which are divided into 2 categories called vague opinion words and clear opinion words. These identify implicit features and cluster features. Feature clustering depends on corresponding opinion words, features similarity and feature structures. Context information enhances clustering in procedure, which is useful in clustering. Results proved that the method performed very well.

Peñalver-Martinez *et al*. (2014) proposed an innovative opinion mining method that took advantage of new Semantic Web-guided solutions to enhance results from conventional natural language processing techniques and sentiment analysis processes. The new methodology aims include:

- Improving feature-based opinion mining using ontologies at the feature selection stage
- Providing sentiment analysis with a new vector analysis-based method. The method was tested on a real-world movie review-themed scenario resulting in promising results compared to conventional approaches.

Khan *et al*. (2014) presented an algorithm for a hybrid approach based twitter feeds classification which included pre-processing steps before feeding text to a classifier. Results show that the new technique overcomes the earlier limitations in achieving higher accuracy compared to other techniques.

Vinodhini and Chandrasekaran (2014) introduced two hybrid models, one PCA with bagging and the other with Bayesian boosting for opinion classification of product reviews. Results were compared to 2 individual statistical learning based classifier models i.e., logistic regression and Support Vector Machine (SVM). This study found that hybrid methods yielded better results regarding 4 quality measures including correctness, misclassification rate, completeness and effectiveness in classifying opinion as positive/negative.

METHODOLOGY

This study investigates product reviews opinion mining. For feature extraction TDF x IDF is used. Feature selection is achieved through MI, IG, FS and PCA.

The methodology flowchart followed is seen in Fig. 1:

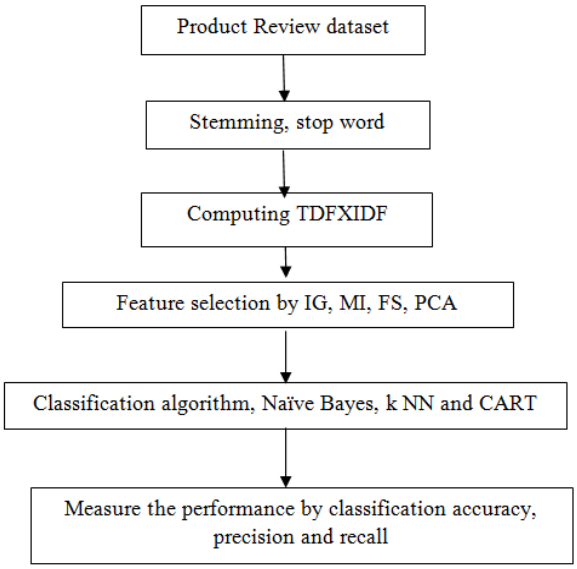

Fig. 1: Flowchart of the methodology

Dataset:

Experiments were carried with customer reviews of 8 products: Two digital cameras, a DVD player, one MP3 player, two cellular phones, one router and anti-virus software. Reviews of the first five products are a benchmark data set from (www.cs.uic.edu). All reviews are from amazon.com. Products on these sites have many reviews each having a text review and a title.

Pre-processing:

Stop words: Some common words are not informative and thus are called stop words. The strategy to determine a stop list is to sort terms by collection frequency (total times a term appears in document collection) and take the most frequent terms (stop words). These words are discarded during indexing.

Stop words do not carry information from a non-linguistic view and are used to remove non-information bearing words from documents and to lower noise. Sentence explanation is held after stop-words removal. To organize a large corpus, removing stop words gives advantages like saving space, helping deduce noises and keeping core words, to ensure efficient processing later.

Stemming: When tokens are created, the next step is converting them to standard form, a process called stemming/lemmatization. Its advantage is that it reduces distinct types in text corpus and increases occurrence frequency of individual types (Maas *et al*., 2011).

Stemming is where different morphological word variants are mapped into their base or common word which is named stem. Present day indexing and search systems support word stemming. Stemming is done by removing attached suffixes/prefixes from index terms prior to actual term assignment to index. Term stem represents a broader concept and so stemming increases retrieved documents in IR systems. The 'stem' in stemming is obtained after application of a rules set without bothering about Part Of Speech (POS) or word occurrence context.

Feature extraction:

TDFxIDF computation: A simple approach is assigning a weight equal to the number of occurrences of term *t* in document *d*. This is called term frequency and denoted tf*t*, *d*, with subscripts denoting term and document in order. But, it is common to use document frequency df*t*, defined as number of documents in a collection using the term *t*.

The Inverse Document Frequency (IDF) of term t is as follows:

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t}. \quad (1)$$

Thus a rare term's IDF is high, while that for a frequent term is low.

Term frequency and IDF together produce a composite weight for every term in a document. The tf-

idf weighting scheme assigns term ta weight in document d given by:

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t. \quad (2)$$

In TF-IDF, term weight based on its frequency in a document is increased; and decreased based on its frequency across document set. To measure such effects, a document's TF is defined as:

$$tf_{t,d} = \frac{number\ of\ occurence\ of\ the\ term\ t}{number\ of\ terms\ in\ the\ document\ d} \quad (3)$$

And the IDF as:

$$idf_t = \log \frac{N}{df_t}. \quad (4)$$

By multiplying TF-IDF weight is defined for term t in document d:

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t. \quad (5)$$

Feature selection: Feature selection is a Research and Development area in machine learning, data mining and statistical pattern recognition. Feature selection chooses an original features subset applied in data mining areas like classification, clustering, association rules and regression. In statistics, feature selection is called subset/variable selection. A feature selection algorithm is ad hoc by nature, but it does not mean that other methods do not exist. It is a fundamental requirement for total search of feature subsets theoretically; as optimal feature selection for supervised learning problems improve performance.

Information Gain (IG): Information Gain (IG) is a popular feature selection method. Using IG reduces noise due to irrelevant features which influence the classifier. IG measures information in bits about class prediction, if information available is presence of a feature and related class distribution.

IG selects test attribute at every tree node. An attribute with highest IG (greatest entropy reduction) is selected as test attribute for current node and it reduces information needed to classify resulting partitions samples.

Entropy measures system disorder/uncertainty. In a classification setting, higher entropy (more disorder) corresponds to a sample with a mixed label collection. Lower entropy corresponds to cases where it contains mostly pure partitions. Entropy of sample D is as follows, in information theory:

$$H(D) = -\sum_{i=1}^{k} P(c_i \mid D) \log_2 P(c_i \mid D) \quad (6)$$

where, $P(c_i \mid D)$ is probability of a data point in D, labeled with class c_i and k is number of classes. $P(c_i \mid D)$ can be estimated from the data as follows:

$$P(c_i \mid D) = \frac{|\{x_j \in D \mid x_j \text{ has label } y_j = c_i\}|}{|D|} \quad (7)$$

Also the weighted entropy of a decision/split are defined as:

$$H(D_L, D_R) = \frac{|D_L|}{|D|} H(D_L) + \frac{|D_R|}{|D|} H(D_R) \quad (8)$$

where, D is partitioned into D_L and D_R because of some split decision. Finally, the IG for a given split can be defined as:

$$Gain(D, D_L, D_R) = H(D) - H(D_L, D_R) \quad (9)$$

Gain is expected reduction in entropy due to knowing an attribute value.

Mutual Information (MI): "Mutual information" is commonly used in word associations and related applications statistical language modeling. It should be termed "point-wise mutual information" as it is not applied to 2 random variables, as in information theory, "mutual information" refers to 2 random variables. This information theory measure compares overall agreement degree between classification and clustering with preference for the latter with high purity (more homogeneous according to classification):

$$MI = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{|C_i \cap P_j| \log(n |C_i \cap P_j|)}{nk^2 |C_i| |P_j|} \quad (10)$$

Principal Component Analysis (PCA): A statistical technique, Principal Component Analysis (PCA) reduces data dimensionality by converting original attribute space. Computing original attributes covariance matrix and extracting eigenvectors, lead to transformed attributes formation. The former (principal components) defines original attribute space from linear transformation to new space where uncorrelated attributes exist. Eigenvectors based on original data variations which they account for are ranked. Usually, the first few transformed attributes account for the most retained data variation while others are rejected. PCA needs no supervision as it does not use class attribute. PCA feature extraction has new attributes using original attributes linear combination and achieving dimensionality reduction by holding on to the highest variance components. Principal components are less than/equal to original variables in numbers. Delineated transformation ensures that a first principal component

has highest variance, leading to great data variability. Hence, every successive component has the highest variance possible under orthogonal constraints-uncorrelated to-that precede components (Isabella and Suresh, 2013).

Dimensions are reduced using PCA when input dimensions are large and components highly correlated PCA calculates an artificial variables smaller set representing observed variable's variance for a variable set. Artificial variables calculated are principal components used as predictor, criterion variable in analysis. PCA orthogonalises variables with resulting in principal components with large variation eliminating components with minimum variation from datasets. PCA observes the following steps when applied on a dataset:

- Mean subtracted from every data dimensions produce a zero mean data set
- Calculate Covariance matrix
- Covariance matrix of eigenvectors and eigenvalues are calculated
- Highest eigenvalues are dataset's principal components. Remove less significant eigenvalues to form feature vector
- Derive a new dataset.

Generally, PCA uses vector space transform to reduce large data sets dimensionality. Using mathematical projection, the original data set, which involved many variables, is now interpreted in just a few variables. Hence, an examination of reduced dimension data set permits users to spot data trends, patterns and outlier more easily than possible without performing PCA.
PCA goals are:

- Extracting most important information from data table
- Compressing data set size by keeping important information
- Simplifying dataset description and
- Analysing structure of observations/variables

PCA originated in multivariate data analysis, but, it has wider applications. PCA is usually used in de-noising signals, separating blind source and compressing data.

Fisher Score (FS): FS is a filter based, supervised feature selection method. The most relevant features for classification are determined using the FS. The FS is a supervised method where features with best discriminant ability and class labels are found. If n_i is number of samples in class i, μ_r^i and $(\sigma_r^i)^2$ is mean and variance of class i, (i = 1,…, c) for feature r, then FS is computed as follows:

$$F_r = \frac{\sum_{i=1}^{c} n_i \left(\mu_r^i - \mu_r\right)^2}{\sum_{i=1}^{c} n_i \left(\sigma_r^i\right)^2} \quad (11)$$

Classifiers:

Naïve Bayes: Naive Bayes is a simple/effective classification algorithm used for document classification (Melville *et al*., 2009; Xia *et al*., 2011; Zhang *et al*., 2011) whose idea is to estimate categories probabilities in a test document through use of joint probabilities of words/categories. The model's naïve part assumes word independence. The assumption's simplicity makes Naive Bayes classifier computation highly efficient.

Naïve Bayes attribute independence assumption works very well for text categorization at a word feature level. When attributes are large, independence assumption permits an attribute's parameters to be learned separately, thus simplifying learning. There are two event models. Of the two, the multi-variate model uses a document event model, with binary words occurrence being event attributes. Here a model does not consider multiple word occurrences in the same document. But, when multiple word occurrences are meaningful, then a multinomial model is resorted to, where multiple word occurrences a are considered by multinomial distribution.

Bayes theorem based statistical classifiers are Naïve Bayes classifiers using a probabilistic approach to predict data class through data matching to class with highest posterior probability:

$$P\left(C_i|V\right) = \frac{P\left(V|C_i\right)P\left(C_i\right)}{P\left(V\right)} \quad (12)$$

where, $V = \left(v_1, ..., v_n\right)$ is the document represented in n-dimensional attribute vector and c_1, ……, c_m represents m class.

RESULTS AND DISCUSSION

The opinions are collected from amazon website. The experiments were carried out using customer reviews of 8 products: Two digital cameras, one DVD player, one MP3 player, two cellular phones, one router and one anti-virus software. Features are extracted using TDF x IDF and Feature selection is achieved using MI, IG, PCA and FS. The selected features were classified using Naïve Bayes and k Nearest Neighbor (KNN). Results obtained for classification accuracy are listed in Table 1.

It is observed from Table 1 and Fig. 2 that the PCA achieves the best accuracy of 82.53% for classification accuracy and PCA is better by 2.96% when compared

Table 1: Classification accuracy for various feature selection methods

	Classification accuracy	
Feature selection	Naïve Bayes	KNN
MI	0.78	0.7395
IG	0.8011	0.7958
FS	0.8074	0.7958
PCA	0.8253	0.8260

Table 2: Precision and recall for various feature selection methods

	Precision		Recall	
Feature selection	Naïve Bayes	KNN	Naïve Bayes	KNN
MI	0.79	0.74	0.77	0.73
IG	0.81	0.81	0.80	0.80
FS	0.81	0.81	0.81	0.80
PCA	0.83	0.83	0.82	0.82

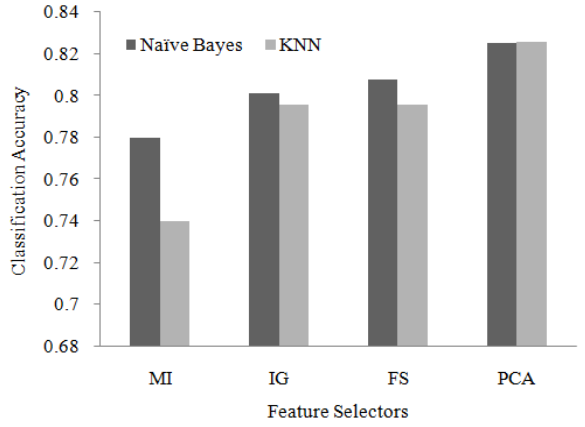

Fig. 2: Classification accuracy obtained for various methods

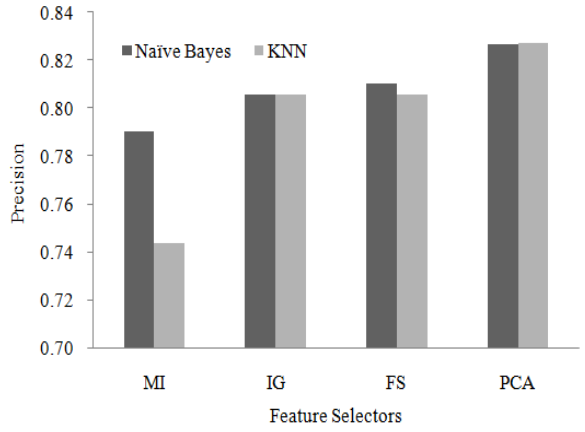

Fig. 3: Precision obtained for various methods

to IG feature selection and by 2.19% when compared to FS feature selection for Naïve Bayes classifier. KNN has lower accuracy when compared to Naïve Bayes for all feature selectors with the exception of PCA. KNN with PCA achieves the highest classification accuracy of 82.6% higher by 0.09% than Naïve Bayes.

It is observed from Table 2 that the PCA achieves the best precision and recall of 0.83and 0.82 for both Naïve Bayes and KNN classifier (Fig. 3 and 4).

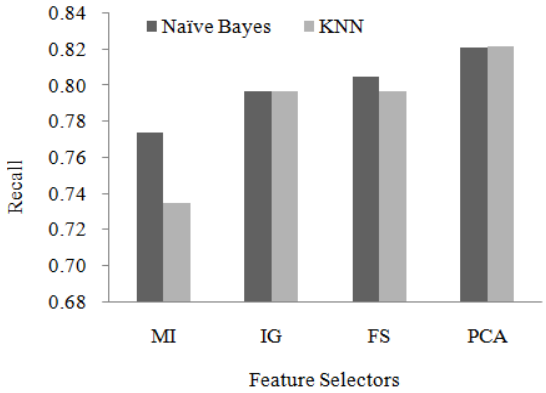

Fig. 4: Recall obtained for various methods

CONCLUSION

A part of information-gathering is finding what people think. The availability of opinion-rich resources like online review sites and personal blogs, chances arise as people use it to understand others opinions. This study investigates feature extraction methods efficacy to classify product reviews. Features from reviews are extracted by using TDF x IDF. Feature selection is through MI, IG, FS and PCA. Naïve Bayes classifier classifies features as positive/neutral/negative. Results prove that the feature selection improves the efficiency of the classifier. The extracted features using PCA achieved the best classification accuracy for both Naïve Bayes and KNN. Further work can be carried out to investigate the optimizing of selecting the feature subsets using swarm intelligence.

REFERENCES

Bollegala, D., D. Weir and J. Carroll, 2013. Cross-domain sentiment classification using a sentiment sensitive thesaurus. IEEE T. Knowl. Data En., 25(8): 1719-1731.

Buche, A., M.B. Chandak and A. Zadgaonkar, 2013. Opinion mining and analysis: A survey. Int. J. Nat. Lang. Comput., 2(3): 39-48.

Chen, J., Y. Liu, G. Zhang, Y. Cai, T. Wang and H. Min, 2013. Sentiment analysis for cantonese opinion mining. Proceeding of the 4th International Conference on Emerging Intelligent Data and Web Technologies (EIDWT, 2013). Xi'an, pp: 496-500.

Cho, K.S., J.S. Ryu, J.H. Jeong, Y.H. Kim and U.M. Kim, 2010. Credibility evaluation and results with leader-weight in opinion mining. Proceeding of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). Huangshan, pp: 5-8.

Ding, J., Z. Le, P. Zhou, G. Wang and W. Shu, 2009. An opinion-tree based flexible opinion mining model. Proceeding of the International Conference on Web Information Systems and Mining (WISM, 2009). Shanghai, pp: 149-152.

Ding, X., B. Liu and P.S. Yu, 2008. A holistic lexicon-based approach to opinion mining. Proceeding of the International Conference on Web Search and Data Mining (WSDM'08), pp: 231-240.

Duan, J. and J. Zeng, 2013. Mining opinion and sentiment for stock return prediction based on web-forum messages. Proceeding of the 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD, 2013), pp: 984-988.

Isabella, J. and R.M. Suresh, 2013. Analysis and evaluation of feature selectors in opinion mining. Indian J. Comput. Sci. Eng., 3(6): 757-762.

Khan, F.H., S. Bashir and U. Qamar, 2014. TOM: Twitter opinion mining framework using hybrid classification scheme. Decis. Support Syst., 57: 245-257.

Liu, L., Z. Lv and H. Wang, 2012. Opinion mining based on feature-level. Proceeding of the 5th International Congress on Image and Signal Processing (CISP, 2012). Chongqing, pp: 1596-1600.

Maas, A.L., R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng and C. Potts, 2011. Learning word vectors for sentiment analysis. Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1: 142-150.

Melville, P., W. Gryc and R.D. Lawrence, 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. Proceeding of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp: 1275-1284.

Pak, A. and P. Paroubek, 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceeding of the 7th International Conference on Language Resources and Evaluation (LREC'10), pp: 1320-1326.

Pang, B. and L. Lee, 2008. Opinion mining and sentiment analysis. Found. Trends Inf. Retrieval, 2(1/2), 1-135.

Peñalver-Martinez, I., F. Garcia-Sanchez, R. Valencia-Garcia, M.Á. Rodríguez-García, V. Moreno, A. Fraga and J.L. Sánchez-Cervantes, 2014. Feature-based opinion mining through ontologies. Expert Syst. Appl., 41(13): 5995-6008.

Vinodhini, G. and R.M. Chandrasekaran, 2012. Sentiment analysis and opinion mining: A survey. Int. J. Adv. Res. Comput. Sci. Softw. Eng., 2(6).

Vinodhini, G. and R.M. Chandrasekaran, 2014. Measuring the quality of hybrid opinion mining model for e-commerce application. Measurement, 55: 101-109.

Wu, H. and X. Gu, 2014. Reducing over-weighting in supervised term weighting for sentiment analysis. Proceeding of COLING 2014, 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, pp: 1322-1330.

Xia, R., C. Zong and S. Li, 2011. Ensemble of feature sets and classification algorithms for sentiment classification. Inform. Sciences, 181(6): 1138-1152.

Zhang, Z., Q. Ye, Z. Zhang and Y. Li, 2011. Sentiment classification of internet restaurant reviews written in Cantonese. Expert Syst. Appl., 38(6): 7674-7682.