International Journal of Engineering & Technology, 7 (1.3) (2018) 140-145

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

An improved wrapper-based feature selection for efficient opinion mining

Prasanna Moorthi N[1]*, Mathivanan V[2]

[1]Research Scholar, Computer Science and Engineering Department
AMET University, Chennai.
[2]Research Supervisor, Computer Science and Engineering Department
AMET University, Chennai
**Email: Prasannamoorthi.n@gmail.com*

Abstract

Opinion mining analyses people's opinions, evaluations, sentiments, attitudes, appraisals and emotions to entities like products, organizations, services, issues, individuals, topics, events and their attributes. It is a large problem space having high feature dimensionality. Feature extraction is important in opinion mining as customers do not usually express product opinions totally, but separately based on individual features. Two tasks should be accomplished in feature-based opinion mining. First, product features on which reviewers expressed opinions must be identified and extracted. Second, opinion orientation or polarities must be determined. Finally, opinion mining summarizes extracted features and opinions. In this work a novel wrapper based feature selection mechanism using concept based feature expansion is proposed. The wrapper based technique uses the principles of evolutionary algorithms.

Keywords: *Opinion Mining; Feature Extraction; Wrapper Based Feature Selection; Concept Based Feature Expansion*

1. Introduction

Generally opinions are expressed on anything, e.g., a product, service, topic, individual, organization, or an event. A general term object denotes the entity commented on. An object has a components (or parts) set and an attributes set. Each component can have sub-components and attributes set etc. Hence, an object is hierarchically decomposed depending on part-of relationship. Beliefs, opinion, emotions and sentiments are private states parts which cannot be observed. These are expressed in documents through subjective words which identify private states using specific dictionaries like WordNet or SentiWordNet. Opinion mining analyses customer's opinions using product reviews providing information including opinions polarity.

Preprocessing [1] is often seen as a fundamental step for Sentiment Analysis, but rarely is it carefully evaluated, thus leaving the open question of why and to what extent does it increase the accuracy of the classifier. Stemming techniques put word variations like \great", \greatly", \greatest", and \greater" all into one bucket, effectively decreasing entropy and increasing the relevance of the concept of \great". In other words, Stemming allows us to consider in the same way nouns, verbs and adverbs that have the same radix. Stop words are words which are filtered out in the preprocessing step. These words are, for example, pronouns, articles, etc. It is important to avoid having these words within the classifier model, because they can lead to a less accurate classification.

Feature based sentiment analysis include feature extraction, sentiment prediction, sentiment classification and optional summarization modules. Feature extraction identifies those product aspects which are being commented by customers, sentiment prediction identifies the text containing sentiment or opinion by deciding sentiment polarity as positive, negative or neutral and finally summarization module aggregates the results obtained from previous two steps. Feature extraction process takes text as input and generates the extracted features in any of the forms like Lexico-Syntactic or Stylistic, Syntactic and Discourse based [2]. Feature extraction in sentiment analysis is facing different issues like large feature space problems, redundancy, domain dependency, difficulty in implicit feature identification.

Feature selection is the process of searching for a feature subset from the original features which is adequate to perform the classification task. Feature selection is able to eliminate redundant features; thus, it assists to improve classification accuracy. Moreover, feature selection is able to reduce the complexity of the learned classifier; consequently, it makes the execution of the learned classifier faster. Wrapper methods use the predictor as a black box and the predictor performance as the objective function to evaluate the variable subset. Since evaluating 2N subsets becomes a NP-hard problem, suboptimal subsets are found by employing search algorithms which find a subset heuristically. A number of search algorithms can be used to find a subset of variables which maximizes the objective function which is the classification performance. The Branch and Bound method used tree structure to evaluate different subsets for the given feature selection number. But the search would grow exponentially for higher number of features. Exhaustive search methods can become computationally intensive for larger datasets. Therefore simplified algorithms such as sequential search or evolutionary algorithms such as Genetic Algorithm (GA) or Particle Swarm Optimization (PSO) which yield local

optimum results are employed which can produce good results and are computationally feasible.

We broadly classify the Wrapper methods [3] into Sequential Selection Algorithms and Heuristic Search Algorithms. The sequential selection algorithms start with an empty set (full set) and add features (remove features) until the maximum objective function is obtained. To speed up the selection, a criteria is chosen which incrementally increases the objective function until the maximum is reached with the minimum number of features. The heuristic search algorithms evaluate different subsets to optimize the objective function. Different subsets are generated either by searching around in a search space or by generating solutions to the optimization problem. Events are occurrence of the feature and occurrence of the class. To test whether the occurrence of a specific feature and the occurrence of a specific class are independent. If the two events are dependent, the occurrence of the feature can be used to predict the occurrence of the class. The features are selected in which the occurrence is highly dependent on the occurrence of the class. When the two events are independent, the observed count is close to the expected count, thus a small chi square score. So a high value of indicates that the hypothesis of independence is incorrect. In other words, the higher value of the score, the more likelihood the feature is correlated with the class, thus it should be selected for model training.

Recently, evolutionary search techniques such as genetic algorithm, genetic programming and Particle Swarm Optimization (PSO) have been used widely to search for feature subsets in the feature selection method. Evaluation methods in feature selection can be divided into the wrapper methods and the filter methods. A wrapper method uses a classifier to evaluate the feature subsets. A wrapper-based feature selection [4] is often computationally intensive since every evaluation of feature subsets requires training a classifier and then testing its performance. In wrapper method, features are selected based on their classification performance using a specific classifier.

Classifying entire documents according to the opinions towards certain objects is called as sentiment classification. One form of opinion mining in product reviews is also to produce feature-based summary. To produce a summary on the features, product features are first identified, and positive and negative opinions on them are aggregated. Features are product attributes, components and other aspects of the product. The effective opinion summary, grouping feature expressions which are domain synonyms is critical. It is very time consuming and tedious for human users to group typically hundreds of feature expressions that can be discovered from text for an opinion mining application into feature categories. Some automated assistance is needed. Opinion summarization does not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as the classic text summarization [5].

2. Literature Survey

Asghar et al [2] discussed on existing techniques and approaches for feature extraction in sentiment analysis and opinion mining. A systematic literature review process was adapted to identify areas well focused by researchers; least addressed areas were also highlighted giving an opportunity to researchers for further work. Most and least commonly used feature selection techniques were identified to find research gaps for future work.

Chandrashekar &Sahin [3] provided an overview of some of the methods present in literature. The proposed method aims to provide a generic introduction to variable elimination which can be applied to a wide array of machine learning problems. Focused on Filter, Wrapper and Embedded methods. Some of the feature selection techniques were discussed on standard datasets to demonstrate the applicability of feature selection techniques.

Tran et al [4] proposed a wrapper-based feature selection method to improve the ability of a classifier able to classify incomplete datasets. In order to achieve the purpose, the feature selection method evaluates feature subsets using a classifier able to classify incomplete datasets. Empirical results on 14 datasets using particle swarm optimization for searching feature subsets and C4.5 for evaluating the feature subsets in the feature selection method show that the wrapper-based feature selection is not only able to improve classification accuracy of the classifier, but also able to reduce the size of trees generated by the classifier.

Angiani et al [6] aims to highlight the importance of preprocessing techniques and show how they can improve system accuracy. In particular, some different preprocessing methods are presented and the accuracy of each of them is compared with the others. The purpose of this comparison is to evaluate which techniques are effective. Proposed method presented the reasons why the accuracy improves, by means of a precise analysis of each method.

Samsudin et al [7] used a feature selection technique based on artificial immune system to select the appropriated features for opinion mining. Experiments with 2000 online movie reviews illustrated that the technique has reduced 90% of the features and improved opinion mining accuracy up to 15% with k Nearest Neighbor classifier and upto 6% with Naïve Baiyes classifier.

Kumar & Abirami [8] performed an experimental study for different feature extraction or selection techniques available for opinion mining task. This experimental study is carried out in four stages. First, the data collection process has been done from readily available sources. Second, the pre-processing techniques were applied automatically using the tools to extract the terms, POS (Parts-of-Speech). Third, different feature selection or extraction techniques were applied over the content. Finally, the empirical study was carried out for analyzing the sentiment polarity with different features.

Recently, a move from traditional word-based approaches to concept-based approaches has started. Schouten & Frasincar [9] showed by using a simple machine learning baseline, that concepts are useful as features within a machine learning framework. Experiments show that the performance increases while including the concept based features**.**

Shang et al [10] proposed a feature selection method called fitness proportionate selection Binary Particle Swarm Optimization (F-BPSO). Binary Particle Swarm Optimization (BPSO) is the binary version of particle swam optimization and can be applied to feature selection domain. F-BPSO is a modification of BPSO and can overcome the problems of traditional BPSO including unreasonable update formula of velocity and lack of evaluation on every single feature. Then, some detailed changes are made on the original F-BPSO including using fitness sum instead of average fitness in the fitness proportionate selection step. The modified method is, thus, called fitness sum proportionate selection binary particle swarm optimization (FS-BPSO). Moreover, further modifications are made on the FS-BPSO method to make it more suitable for sentiment classification oriented feature selection domain. The modified method is named as SCO-FS-BPSO where SCO stands for “sentiment classification-oriented”. Experimental results show that in benchmark datasets original F-BPSO is superior to traditional BPSO in feature selection performance and FS-BPSO outperforms original F-BPSO. Besides, in sentiment classification domain, SCO-FS-BPSO which is modified specially for sentiment classification is superior to traditional feature selection methods on subjective consumer review datasets.

3. Methodology

In this section detail of feature extraction, selection and classification is performed.

3.1 Cell Phones and Accessories dataset

To evaluate the strength of our method at capturing fashion dynamics, are interested in real-world datasets that (a) are broad enough to capture the general tastes of the public, and (b) temporally span a long period so that there are discernibly different visual decision factors at play during different times. The two datasets used are from Amazon.com. Consider two large categories that naturally encode fashion dynamics (within the U.S.) over the past decade, namely Women's and Men's Clothing & Accessories, each consisting of a comprehensive vocabulary of clothing items. The images available from this dataset are of high quality (typically centered on a white background) and have previously been shown to be effective for recommendation tasks (though different from the one consider here). Processed each dataset by taking users' review histories as implicit feedback and extracting visual features fi from one image of each item i. Discarded users u who have performed fewer than 5 actions, i.e., for whom jI+ u j< 5 [11]

Blogs, review sites and micro blogs provide a good understanding of the reception level of products and services.

- **Blogs:** The name associated to universe of all the blog sites is called blogosphere. People write about the topics they want to share with others on a blog. Blogging is a happening thing because of its ease and simplicity of creating blog posts, its free form and unedited nature. A large number of posts was on virtually every topic of interest on blogosphere. Sources of opinion in many of the studies related to sentiment analysis, blogs are used.
- **Review Sites:** Opinions are the decision makes for any user in making a purchase. The user generated reviews for products and services are largely available on internet. The sentiment classification uses reviewer's data collected from the websites like www.gsmarena.com (mobile reviews), www.amazon.com (product reviews), www. CNET-download.com (product reviews), which hosts millions of product reviews by consumers [12].
- **Micro-blogging:** A very popular communication tool among Internet users is micro-blogging. Millions of messages appear daily in popular web-sites for micro-blogging such as Twitter, Tumblr, Facebook. Twitter messages sometimes express opinions which are used as data source for classifying sentiment.

Term Frequency -Inverse Document Frequency (TF-IDF)

A measure of how significant a term is in a document collection is given by the term frequency. In a document collection term frequency of a term (t_i) is defined as in equation (1):

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

where $n_{i,j}$ is the number of occurrences of the considered as term (t_i) in document d_j, and the denominator is the sum ofthe number of occurrences of all terms in document d_j. The inverse document frequency is a measure of the importance of the term in the entire document collection.

$$idf_i = \log \frac{|\,D\,|}{|\{d : t_i \in d\}|} \tag{2}$$

where | D | : total number of documents in the corpus and $|\{d : t_i \in d\}|$: number of documents where the term t_i appears. Then, when inverse document frequency factor is incorporated, the weight of terms that occur very frequently in the collection diminishes and the weight of terms that occur rarely increases.

$$\{tf - idf\}_{i,j} = tf_{i,j} \times idf_i \tag{3}$$

3.2 Particle Swarm Optimization (PSO)

PSO, was first proposed by Kennedy and Eberhart (1995). It was inspired by the social behavior of birds flocking or fish schooling. A swarm has some particles, each particle has a position component representing a specific solution, and a velocity component representing the direction of a particle's movement in the solution space. PSO is an iterative optimization algorithm with three main steps. The first step is to initialize the population by generating each particle's velocity component and position component randomly. The second step is to evaluate solutions represented by particles' positions. The final step is to update particles' velocities and then update particles' positions using the following formulas. The second and third steps are repeated until the stop criterion is met.

$$v_i^{t+1} = w * v_i^t + c_1 * rand * (pbest - x_i^t) + c_2 * rand * (\mathrm{g}best - x_i^t)$$
$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{4}$$

In Equation (4), v_i^{t+1} and x_i^{t+1} represent the velocity component and the position component of particle *p* at the*(t* + 1)th iteration, respectively; c_1 and c_2 are confidence coefficients, *rand* is a uniformly distributed random variable ranging from 0 to 1, *w* is the inertia weight. *pbest* means the position of particle *p*'s personal best while *gbest* means the position of all particles' global best.

The right side of Eq. (4) can be divided into three parts. $w * v_i^t$ represents the previous direction, $c_1 * rand * (pbest - x_i^t)$ represents the tendency of moving towards a particle's personalbest, $c_2 * rand * (\mathrm{g}best - x_i^t)$ represents the tendency of moving towards the swarm's global best. The three parts together guide a particle's movement.

Levy Flight Search

Lévy flights comprise sequences of straight-line movements with random orientations. Lévy flights are considered to be 'scale-free' since the straight line movements have no characteristic scale. The distribution of the straight line movement lengths, *l* have a power-law tail [13] as in equation (5):

$$P(l) \sim l^{-\mu} \tag{5}$$

where $1 < \mu < 3$.The sum of the a set $\{l_i\}$ converge to the Lévy distribution, which has the following probability density as in equation (6):

$$L_{\alpha,\gamma}(l) = \frac{1}{\pi}\int_0^{\infty} e^{-\gamma q^{\alpha}} \cos(ql)dq \quad (6)$$

Where α and γ are two parameters that control the sharpness of the graph and the scale unit of the distribution, respectively. The two satisfy $1 < \alpha < 2$and $\gamma > 0$. For $\alpha \to 1$, the distribution becomes Cauchy distribution and for $\alpha \to 2$, the distribution becomes Gaussian distribution. Without losing generality, set the scaling factor γ = 1. Therefore, it can re-scaled $L_{\alpha,1}$ with some constant $b: L_{\alpha,1}(bl) = bL_{\alpha,1}(l)$, for $b \in \Re$.

Since, the analytic form of the Lévy distribution is unknown for generalα, in order to generate Lévy random number, adopted a fast algorithm. Firstly, Two independent random variables *x* and *y* from Gaussian distribution are used to perform a non-linear transformation as in equation (7):

$$v = \frac{x}{|y|^{1/\alpha}} \quad (7)$$

Then the random variable, now in the Lévy distribution, *r*1, is generated using the following nonlinear transformation as in equation (8):

$$r_1 = \{(K(\alpha)-1)e^{-\frac{|v|}{C(\alpha)}} + 1\}v \quad (8)$$

where the values of parameters *K(α)* and *C(α)* are given in [13].

3.3 Proposed PSO based Feature selection

We focus on [10] binary sentiment classification, where each document is classified either as “positive” or “negative” according to author’s tendency of opinion. In the classification process, every document in the dataset is transformed under a specific model and represented as a feature vector during data preprocessing. One commonly used model for document representation is unigram bag-of-words model (BoW). UnderBo W model, number of dimensions of each feature vector is the number of different words in the whole text dataset. The vector assigns “1” to *d*th dimension if the text contains corresponding word, and assigns “0” if it does not.

BPSO can be used to perform feature selection. Similar to the genetic algorithm-based method, a possible feature subset is represented as a binary vector in BPSO-based method. The length of vector is equal to the number of the available features and the value of each bit represents whether the corresponding feature is selected or not. The main aim of feature selection is to increase the classification accuracy on the dataset. Therefore, when evaluating the fitness value of a particle, the training set will be equally divided into 10 folds and cross-validation will be run on the training set with supervised machine learning scheme using the feature subset the particle represents. Then the particle’s fitness value will be the average accuracy of all 10runs.

PSO based feature selection
initialize parameters of PSO
randomly initialize swarm
while stopping criterion not met do
 for i=1 to swarmsize do
  calculate P_i's fitness value
  update the pbest of P_i
  update the gbest of P_i
 end
 for i=1to swarmsize do
  for j=1 to dimension do
   update the velocity of P_i
   update the position of P_i
  end
 end
end
return the best feature subset found by the swarm

3.4 Classifiers

3.4.1 Naïve Bayesian Classifications

Naïve Bayesian method is one of the popular techniques for text classification. It has been shown to perform extremely well in practice by many researchers. Given a set of training documents D, each document is considered an ordered list of words. Let $w_{d_i,k}$ denotes the word in position k of document d_i , where each word is from the vocabulary $V = < w_1, w_2, ..., w_{|v|} >$,where vocabulary is the set of all words considered for classification, and let a set of pre-defined classes be $C = < c_1, c_2 \ldots, c_{|c|} >$. In order to perform classification, it is need to compute the posterior probability, $P[c_j \mid d_i]$. Based on the Bayesian probability and the multinomial model as in equation (9):

$$P[c_j] = \sum_i P[c_j / d_i] / D \quad (9)$$

To eliminate zeros, Laplacian smoothing can be used [14], which simply adds one to each count:

$$P[w_t / c_j] = \frac{1 + \sum_{i-1}^{D} N(w_t, d_i) P(c_j / d_i)}{V + \sum_{s-1}^{V}\sum_{i-1}^{D} N(w_s, d_i) P(c_j / d_i)}$$

where $N(w_s, d_i)$ is the number of times the word, w_t, occurs in document, di, and $P(c_j / d_i)$ {0,1} depends on the class label of the document. Finally, assuming that the probabilities of the words are Independent given the class as in equation (10):

$$P[w_t / c_j] = \frac{P[c_j]\prod_{k-1}^{d_i} P[w_{d_{i.k}} / c_j]}{\sum_{r-1}^{c} P[c_r]\prod_{k-1}^{d_i} P[w_{d_{i.k}} / c_j]} \quad (10)$$

In the naïve Bayes classifier, the class with the highest $P[c_j / d_i]$ is assigned as the class of the document. Thus it is a supervised learning method. A Bayesian Classifier is a simplest

probabilistic classifier based on Bayes theorem. In text classification, to determine the most probable class or group, a document falls into, Bayes rule is used.

3.4.2 K-nearest neighbor (KNN)

KNN is a simple machine learning algorithm. In this algorithm, the objects are classified based on the majority of its neighbor. The class assigned to the object is most among its k nearest neighbors. The KNN classification algorithm classifies the instances or objects based on their similarities to instances in the training data. In KNN, selection is based on majority voting or distance weighted voting. KNN is unsupervised text classification algorithm and it works efficiently when the training set is large. Consider the vector A and set of M labeled instances $\{a_i, b_i\}$. The classifier predicts the class label of A on the predefined N classes. The KNN classification algorithm finds the k nearest neighbors of A and determines the class label of A using majority vote. KNN classifier applies Euclidean distances as the distance metric.

$$Dist(X,Y) = \sqrt{\sum (X_i - Y_i)^2} \quad (11)$$

4. Results and Discussion

Table 1 and figure 1 to 2 shows the results of true positive rate and precision respectively

Table 1: True Positive Rate and Precision for Levy PSO-NB

	Chi Square-KNN	Chi Square-NB	PSO-KNN	PSO-NB	Levy PSO-KNN	LevyP-SO-NB
True Positive Rate	0.878467	0.8889	0.913633	0.9303	0.945	0.955733
Precision	0.876533	0.8865	0.911733	0.9288	0.94356667	0.954933

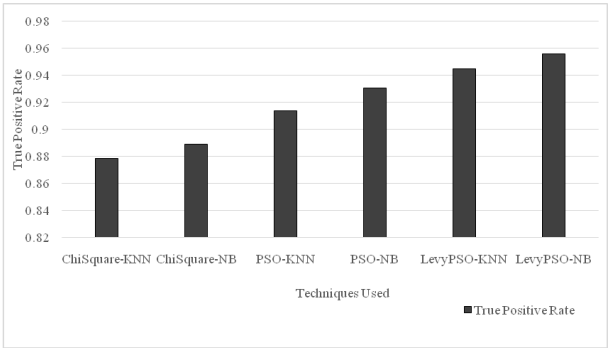

Fig. 1: True Positive Rate for Levy PSO-NB

Table 1 and Figure 1 shows that the true positive rate for Levy PSO-NB performs better by 8.43% than ChiSquare-KNN, by 7.25% than ChiSquare-NB, by 4.5% than PSO-KNN, by 2.7% than PSO-NB andby 1.13% than LevyPSO-KNN.

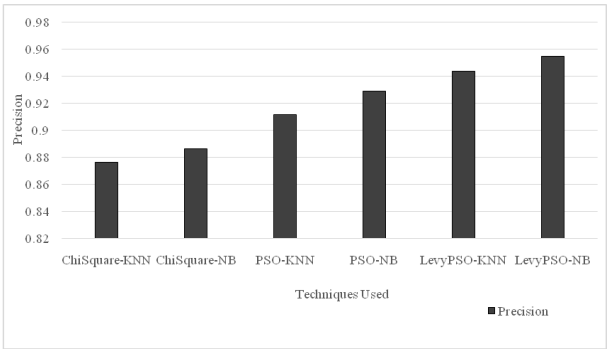

Fig.2: Precision for Levy PSO-NB

Table 1 and Figure 2 shows that the precision for Levy PSO-NB performs better by 8.56% than ChiSquare-KNN, by 7.43% than ChiSquare-NB, by 4.63% than PSO-KNN, by 2.77% than PSO-NB and by 1.19% than LevyPSO-KNN.

5. Conclusion

Opinion mining is used to analyze the sentiments expressed by people on the web through reviews. In recent years, large attention has been given to opinion mining because of its wide range of possible applications. Feature selection methods provide a criterion for eliminating terms from document corpus to reduce vocabulary space. Results show that the true positive rate for Levy PSO-NB performs better by 8.43% than ChiSquare-KNN, by 7.25% than ChiSquare-NB, by 4.5% than PSO-KNN, by 2.7% than PSO-NB and by 1.13% than LevyPSO-KNN. Also the precision for Levy PSO-NB performs better by 8.56% than ChiSquare-KNN, by 7.43% than ChiSquare-NB, by 4.63% than PSO-KNN, by 2.77% than PSO-NB and by 1.19% than LevyPSO-KNN.

References

[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., &Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media (pp. 30-38). Association for Computational Linguistics.

[2] Asghar, M. Z., Khan, A., Ahmad, S., &Kundi, F. M. (2014). A review of feature extraction in sentiment analysis. Journal of Basic and Applied Scientific Research, 4(3), 181-186.

[3] Chandrashekar, G., &Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.

[4] Tran, C. T., Zhang, M., Andreae, P., &Xue, B. (2016). Improving performance for classification with incomplete data using wrapper-based feature selection. Evolutionary Intelligence, 9(3), 81-94.

[5] Buche, A., Chandak, D., &Zadgaonkar, A. (2013). Opinion mining and analysis: a survey. arXiv preprint arXiv:1307.3336.

[6] Angiani, G., Ferrari, L., Fontanini, T., Fornacciari, P., Iotti, E., Magliani, F., &Manicardi, S. (2016). A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. In KDWeb.

[7] Samsudin, N., Puteh, M., Hamdan, A. R., &Nazri, M. Z. A. (2013, July). Immune based feature selection for opinion mining. In Proceedings of the World Congress on Engineering (Vol. 3, pp. 3-5).

[8] Kumar, J. A., &Abirami, S. (2015). An Experimental Study Of Feature Extraction Techniques In Opinion Mining. International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), 4(1).

[9] Schouten, K., &Frasincar, F. (2015, May). The benefit of concept-based features for sentiment analysis. In Semantic Web Evaluation Challenge (pp. 223-233). Springer, Cham.

[10] Shang, L., Zhou, Z., & Liu, X. (2016). Particle swarm optimization-based feature selection in sentiment classification. Soft Computing, 20(10), 3821-3834.

[11] He, R., &McAuley, J. (2016, April). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In Proceedings of the 25th International Conference on World Wide Web (pp. 507-517). International World Wide Web Conferences Steering Committee.
[12] G.Vinodhini and RM.Chandrasekaran, Sentiment Analysis and Opinion Mining: A Survey, Volume 2, Issue 6, June 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.
[13] He, S. (2010). Training Artificial Neural Networks Using Lévy Group Search Optimizer. Multiple-Valued Logic and Soft Computing, 16(6), 527-545.
[14] Qing Cao, Mark A. Thompson, Yang Yu, (2012), "Sentiment analysis in decision sciences research: An illustration to IT governance, Decision Support Systems, http://dx.doi.org/10.1016/j.dss.2012.10.026.