Journal of Scientific & Industrial Research
Vol. 75, April 2016, pp. 221-224

An Efficient Feature Selection Technique of Unsupervised Learning Approach for Analyzing Web Opinions

M S Valli[1]* and G T Arasu[2]

[*1]Department of CSE, PSV College of Engineering & Technology, Krishnagiri, Tamil Nadu, India.
[2]AVS Technical Campus, Salem, Tamil Nadu, India.

Received 22 September 2014; revised 14 October 2015; accepted 3 January 2016

Examination of developing Web opinions is probably valuable for realize enduring topics of public like crime and terrorist attack detection, Some participants or users are using the web forum to publicize their thought about particular incident in the world such as committing crime. How the topics are progressed together with the social interaction between contributors and recognize important and influence of participants discussions of various topics through social media. Participants are usually considering opinions to be articulated through adjectives, and make wide use of moreover general dictionaries or specialist to provide the appropriate adjectives. However, analyzing and clustering of Web opinions is really difficult. Unlike the documents of Web opinions are tiny and sparse with noisy text content, typical Web opinions document clustering method produce unsatisfactory performance. Feature selection (FS) is a procedure which efforts to select more informative features. Some Web opinions documents have too many repeated and irrelevant texts for classification or clustering. Feature selection method can recover this problem instead of classification and clustering algorithms. The main aspect of feature selection is give high accuracy performance with minimal feature subset. In this paper, we propose the unsupervised rough set method for clustering text for Web opinion mining. We conducted more experiments and had benchmarked with the unsupervised algorithm which gives higher micro accuracy results.

Keywords: Feature Based Summarization, Feature Selection, Support Vector Machine, Rough Set Theory.

Introduction

Nowadays, Opinion mining is an imperative research of web data mining field [1, 2, 3]. The intention of opinion mining is to evaluate the sentiments thoughts, feedback and comments expressed by people on the web. It gives the large concentration in recent years because of its wide range of possible applications used in internet and most of users have been using it for communicating with others for sharing ideas[4,5,6]. Most of enterprises and corporate people can make use of opinion mining performance for marketing and product manufacturing researches[7]. Conditionally on the level of interest, opinion mining has two levels such as Document-level Opinion Mining and Feature-based Summarization (FBS). FBS is used to identify the target features mentioned in the examination and find out sentiment expressed by the author. A Web forum is an effective platform for communicating personal and public opinions, experiences, comments, and thoughts in discussion threads[8]. Web users are capable to share their opinions about personal and social related to friends whoever is connected in their social network[9,10,11]. In the opinion extraction task, on the other hand, it is unclear what should be extracted, since the opinions include subjective expressions on various topics. Previous work does not sufficiently discuss how customer reviews are reported in web documents and can be structures. For example, to an extreme, the Gray Web Forum in the recent years has focused on topics that might potentially state and encourage biased, offensive, or disruptive behaviors and might disturb the society, or threaten the public or even national safety[12,13,14]. In this paper, we present rough set based unsupervised feature selection techniques in Web opinion mining for clustering information based on conditional attributes, which are components of Web opinion analysis and understanding[13]. Applying the unsupervised techniques K-Means and MRF feature selection to review feedbacks of the customers of Thai restaurant on their products and services[15]. This paper proposed a method for text preprocessing for crack reviews into words and removing stop words. MRF feature selection is consequently implemented for selecting significant features from a

*Author for correspondence
E-mail: senba1983@gmail.com

huge number of features extracted. K-Means is in used for clustering into reviews like positive and negative wise. This feature selection method can efficiently reduce the more number of features in the data set. In Web opinion clustering, we cannot predict the number of clusters because it changes from time to time and also many Web opinions are noisy so they are not giving to any cluster. In our groundwork studies[14], it is found that more Web opinions are noise. Due to all of these reasons performance is poor in Web opinion clustering techniques when it is applied directly.[16] They are analyzing on feature selection and machine learning algorithms for Malay sentiment classification. This work motivated to apply the feature reduction and selection for improve classifier performance.[17] proposed supervised learning is a binary or multi class classification, it is needed to improve efficiency and avoid over fitting.

Unsupervised Feature Selection

Unsupervised feature selection techniques are based on relative dependency measure using rough set theory (RST). Decision class attribute are frequently unidentified or imperfect; in this condition the unsupervised feature selection method is compete essential responsibility to select features [18].

Relative dependency measure

The unsupervised relative dependency evaluate for an attribute subset is distinct as follows:

$$K_R(\{a\}) = |U/IND(R)| / |U/IND(R \cup \{a\})|, \forall\ a \in A \ldots(1)$$

$$K_R(\{a\}) = K_C(\{a\}) \ \& \ \forall\ X \subset R,\ K_X(\{a\}) = K_C(\{a\}) \ldots(2)$$

In above situation, the decision attribute used in the supervised feature selection, is changed by the conditional attribute 'a', which is to be eliminated from the current reduct set R.

Unsupervised relative reduct (USRR) algorithm

The new USRR algorithm is shown below, the algorithm begin with by considering all of the features included in the dataset.

US relative reduct(C), C is the conditional attributes.

(1) R←C
(2) $\forall_a \in C$
(3) If $(K_{R-\{a\}}(\{a\}) == 1)$
(4) R←R-{a}
(5) Return R

In the USRR algorithm each feature is observed iteratively and the relative dependency measure are determined. If the relative dependency is equal to 1 then that feature can be eliminated. These processes persist until all features have been examined.

Methodology

System Architecture Design

In the Proposed system structural design, the collected set of reviews is preprocessed and the words are selected. The stop words are removed at the stage of preprocessing. The different phases of design process are as follows:

Data Preparation

1000 positive feedbacks and 1000 negative feedbacks related to social activities were collected randomly from online forums such as Face book, Twitter, web forums, Linkedin, Google + and some of web forums. More than a few groundwork's had been executed previously to the opinion mining. The online data was standardizing using some methods. As well as there is no meaningful words was removed. The words are converted into the lower case format. These actions were conducted to remove unnecessary words and thus reducing the number of features. In English, languages play a most important role in articulating negative sentiment. The word that has explained exists frequently in the positive and negative classes.

Data Representation

The previous method that is recommended by applying artificial techniques in problem solving is mapping the concept between the Immune System projects. Table 1 shows the mapping between Immune System and FS-INS (Feature Selection based on Immune Network System).

Feature Extraction Step

Our technique combines sentences which are classified as relevant by our RST and do not belong to any cluster in DBSCAN clustering. The input parameters of the algorithm are the set of all type of feedback, the calculated classification model, and

Table 1—Mapping concept of Immune System and FS-INS

Immune System	This Project
Antibody	Features in the population
Antigen	Another feature
Memory of antibody	A group of features which had
undergone	The training phase.

the calculated clustering model (here, it is only important, if a sentence is identified as noise or has any cluster id). The algorithm (cf. algorithm 1) filters out sentences which are not relevant or belong to a cluster, and combines the remaining sentences, if they are consecutive.

Feature Extraction – Algorithm 1:

a. Set the attributes to: Sentences S, Classification Model K, Clustering Model C and Result: List of Statements R
b. R is used to create an empty list of statements and C is the conditional attributes.
c. To check the condition If (KR-{a} ({a}) ==1) R←R-{a} and it will Return R.
d. Removing/combining the statements by using,
   if r1.EndOffset + 1 = r2.StartOffset then
   remove r1 and r2 from R;
   add (combine (r1, r2)) in R;

Note: The method combine takes two consecutive statements and appends the second one to the first one. R contains all pi with i ∈ {1... ν − 1} and p_ν are all sentences which are not a part of an element in R.

Experiments and Result

The performance was evaluated by performing the opinion mining process using Weka application. The accuracy of opinion mining using RST as feature selection was compared to the accuracy of few opinion mining techniques without any feature selection. The accuracy value was calculated using equation is

Accuracy= # of Correct Prediction / # of reviews … (3)

Fig. 1 shows the reduction of features in experiments which varied the number of 1000 positive feedbacks and 1000 negative feedbacks related to social activities. The result shows that the number of features was reduced when RST was utilized as the unsupervised feature selection technique. Table 2 shows the results for feature selection of word as relevant or not. As the tables 2 shows, our method needs only limited training data (5% or 0.5%, respectively.) to obtain good results. There is almost no difference between using 15% or 5% on the feedback dataset. The result shows that it is more complicated to identify the relevant sentences in FS-INS method. The feedback dataset contains 3,283 relevant sentences and 11,806 sentences are not relevant. However, the tables show that RST methods achieve better results than FS-INS. Table III shows the performance of FS-INS as feature selection in opinion mining with three common classifiers i.e. Naïve Bayes, k Nearest Neighbor (k=1) and Support Vector Machine (SVM). The result shows the performance of various opinion mining clustering techniques such as KNN, NB, SVM and unsupervised RST feature selection. Fig. 2 and 3 shows the various result performances of RST and FS-INS feature selection techniques. Analyzing the result Our Proposed method Unsupervised RST gives the better performance the FS-INS.

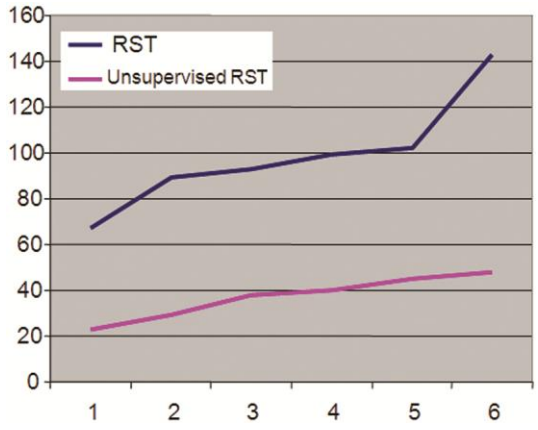

Fig. 1—Reduction of Features

Table 2—Results of the sentence extraction

Method	Data for Method	Accuracy	Not Relevant		Relevant	
			Precision	Recall	Precision	Recall
FS-INS	2.5%	0.6403	0.2591	0.5923	0.8854	0.6503
FS-INS	5%	0.6938	0.8579	0.7556	0.2501	0.3943
FS-INS	10%	0.659	0.8659	0.6969	0.2434	0.4246
FS-INS	15%	0.6525	0.8661	0.6866	0.2443	0.4882
RST	2.5%	0.4918	0.8315	0.4872	0.1694	0.5143
RST	5%	0.8172	0.8912	0.8885	0.4597	0.4667
RST	10%	0.8178	0.8914	0.8892	0.4601	0.4656
RST	15%	0.8173	0.8111	0.8890	0.4479	0.4633

Table 3—Performance of RST, FS-INS in NB, KNN and SMV Classifier

Methods	% Acc with FS-INS
NB	91.04
kNN	79.08
SVM	92.25
RST	93.75

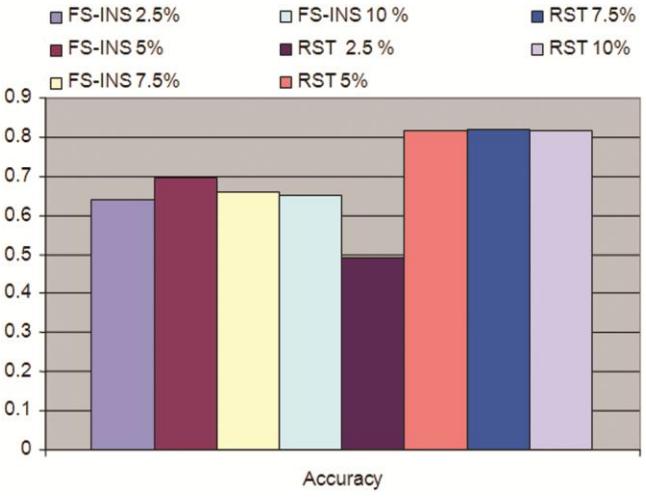

Fig. 2—Accuracy Rate of RST and FS-INS

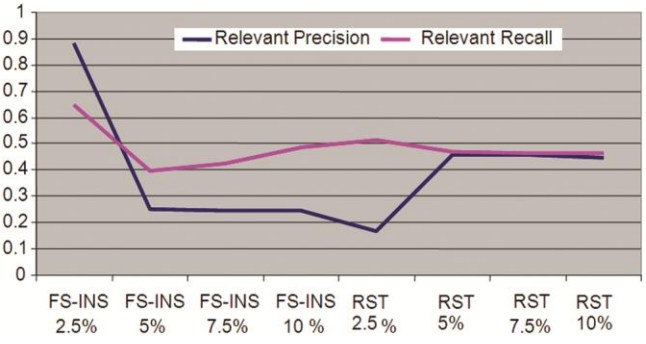

Fig. 3—Relevant Precision and Recall rate of RST and FS-INS

Conclusion

This paper discusses the unsupervised feature selection techniques in opinion mining using feedbacks in social activities in various social networks. Machine leaning approach is a good feature selection method and was proposed earlier. Motivated by the artificial immune system, an unsupervised Rough set feature selection technique was developed in this study. The experiments illustrate the number of features selected by RST was been reduced by 90% and it had triggered the improvement of opinion mining's accuracy up to 15% in FS-INS. The findings point out that the extraction of statements could not only be solved by Text Summarization. On the one hand, this approach can be utilized to help media analysts who could save time by reading news articles and extracting relevant statements.

References

1 Pang B & Lee L, Opinion mining & sentiment analysis, Now Publishers (2008).

2 Branavan S, Chen H, Eisenstein J & Barzilay R, Learning Document-Level Semantic Properties from Free-Text Annotations, *In Proc of the Assoc for Comp Ling* (2008).

3 Jindal N & Liu B, Opinion Spam and Analysis, *Int conf on Web search & Web Data Mining*, (2008) 219-230.

4 Blair-Goldensohn S, Hannan K, McDonald R, Neylon T, Reis G & Reynar J, Building a Sentiment Summarizer for Local Service Reviews, *In WWW 2008 Workshop on NLP Challenges in the Info Explo,* (2008).

5 Lee D, Jeong O & Lee S, Opinion Mining of Customer Feedback Data on the Web, *2nd Int conf on Ubiquitous info mgmt & comm,* (2008) 230-235.

6 Zhang W, Yu C & Meng W, Opinion retrieval from blogs,*conf on info & know Mgmt*, (2007) 831-840.

7 Ghose A & Ipeirotis P, Estimating the Socio-Economic Impact of Product Reviews: Mining Text & Reviewer Characteristics, Info. Sys Research (2008).

8 Yang C C, Ng T D, Wang J, Wei C & Chen H, *Analyzing & visualizing gray Web forum structure, in Proc* Pacific Asia Workshop Intell, Security Info,(2007).

9 Kumar R, Novak J, Raghavan P & Tomkins A, Structure & Evolution of Blog Space, Commun., **47**(2004) 35–39.

10 Mei Q, Liu C, Su H, & Zhai C, A probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs, in Proc *Int WWW Conf, U.K*., (2006).

11 Nardi B A, Schiano D J, Gumbrecht M & Swartz L, Why we blog, Commun. ACM, **47**(2004) 41–46.

12 2 Wang J, Fu T, Lin H & Chen H, A framework for $exploring Gray Web forums: Analysis of Forum-Based comm. In Taiwan, in Proc*. IEEE Int Conf Intell Security Info San Diego,* (2006) 498–503.

13 Zhou Y, Reid E, Qin J, Lai G & Chen H, Domestic Extremist Groups on the Web: Link & Content Analysis, *IEEE Intell Syst*, **20** (2005) 44–51.

14 Yang C C & Ng T D, Terrorism & Crime Related Weblog Social Networks: Link, Content Analysis & Information Visualization, in Proc. *IEEE Int Conf Intell Security Info,* (2007) 55–58.

15 Claypo N, King Mongkut's & Jaiyen S, Opinion mining for Thai Restaurant Reviews using K-Means Clustering & MRF Feature Selection, *IEEE, Know & Smart Tech,*(2015) 28-31.

16 Alsaffar A & Omar N, Study on Feature Selection & Machine Learning Algorithms for Malay Sentiment Classification, *IEEE, in conf Infor Tech & Multi,* (2014) 270 – 275.

17 Baccianella C, Esuli A & Sebastiani F, Feature Selection for Ordinal Text Classification ,conf in Neural Comp, (2014) 557-591.

18 Velayutham C, Rough Set Based Unsupervised Feature Selection Using Relative dependency Measures, *Int J of Comp Intell & Info*, **1**(2011) 120-124.