Kuwait J. Sci. **43 (1)** pp. 129-149, 2016

Pattern and semantic analysis to improve unsupervised techniques for opinion target identification

Khairullah Khan[1,*], Ashraf Ullah[1], Baharum Baharudin [2]

[1]*Institute of Engineering and Computing Sciences, University of Science & Technology Bannu, Khyber Pakhtoonkhawa, Pakistan.*

email: khairullah_k@yahoo.com , ashrafbth@gmail.com

[2]*Department of Computer and Information Sciences, Universiti Teknologi, PETRONAS Malaysia.*

**Corresponding Author: Email: khairullah_k@yahoo.com*

Abstract

This research employs patterns and semantic analysis to improve the existing unsupervised opinion targets extraction technique. Two steps are employed to identify opinion targets: candidate selection and opinion targets selection. For candidate selection; a combined lexical based syntactic pattern is identified. For opinion targets selection, a hybrid approach that combines the existing likelihood ratio test technique with semantic based relatedness is proposed. The existing approach basically extracts frequently observed targets in text. However, analysis shows that not all target features occur frequently in the texts. Hence the hybrid technique is proposed to extract both frequent and infrequent targets. The proposed algorithm employs incremental approach to improve the performance of existing unsupervised mining of features by extracting infrequent features through semantic relatedness with frequent features based on lexical dictionary. Empirical results show that the hybrid technique with combined patterns outperforms the existing techniques.

Keywords: Information retrieval; machine learning; natural language processing; opinion mining; text mining.

1. Introduction

The focus of this study is opinion target identification for the opinion mining process. The problem of opinion target identification is related to the question: "opinion about what?'. Opinion target identification is essential for opinion mining. For example, the in-depth analysis of every aspect of a product based on consumer opinion is equally important for consumers, merchants and manufacturers. In order to compare the reviews, it is required to automatically identify and extract those features, which are discussed in the reviews. Furthermore, analysis of a product at feature level is more important; e.g. which features of the product are liked and which are disliked

by consumers. Hence, feature mining of products is important for opinion mining and summarization. The task of feature mining provides a base for opinion summarization. There are various problems related to opinion target extraction. Generally speaking, if a system is capable of identifying a target feature in a sentence or document, then it must be able to identify opinionated terms or evaluative expressions in that sentence or document. Thus in order to identify opinion targets at sentence or document level, the system should be able to identify evaluative expressions.

Opinion target identification is basically a classification problem, which is defined as: to classify noun phrase or term as opinion target or not (Khairullah *et al.,* 2013). As mentioned earlier in the background study, there are two widely used classification methods i.e. supervised and unsupervised. The supervised method needs prior knowledge annotated through manual process. Unsupervised classification depends on heuristics procedures and rules, which do not need previous knowledge.

The main focus of this paper is on opinion target identification through unsupervised method. There are two main advantages for unsupervised method over supervised: supervised technique need training data which are manually labeled, while unsupervised do not need hand-crafted training datasets. Supervised techniques are generally domain dependent as training data are manually labeled for specific domain, while unsupervised are domain independent. Our main objective is the identification of domain-independent opinion target.

The two most frequently reported unsupervised approaches for opinion target identification are Association Mining (AM) (Agrawal & Srikant, 1994) and the Likelihood Ratio Test (LRT) approach (Dunning, 1993). However these approaches suffer from some limitations. For example, both of these approaches use threshold value, which depends on frequency of terms. The ideal value for threshold is difficult to identify, hence features with low frequencies are misclassified. Some terms with high frequency do not relate to the topic and may not be a feature. These techniques are progressively improved as described in the related work. However, the results are still affected by the threshold values. Keeping in view the limitations of the existing approaches, a novel Semantic Based Likelihood Ratio Test (SLRT) approach is proposed, which combines the LRT with semantic based similarity scoring. The LRT approach has several advantages over association mining approach (Ferreira *et al.,* 2008). Therefore, this work employs likelihood ratio test in hybrid with the semantic based relevance scoring. However, the LRT has certain limitations too. For example, it cannot detect rarely occurred opinion targets. Semantic relatedness is employed for detection of infrequent features to improve the performance of the likelihood relevance scoring. The idea behind this technique is to identify infrequent features through semantic relatedness among the frequent features and infrequent features.

The proposed hybrid technique has two steps. In the first step, LRT is applied to extract frequently occurred opinion targets, while in the second step semantic based relation is applied to identify rarely occurred opinion targets. Answer to question; that why semantic based relation is employed to extract infrequent features; it depends on observations and experiments. This technique assumes that the rarely occurred opinion targets have high chance of relation with the frequently occurred features. For example, picture, image and photo are closely related; hence when either of the features occurs frequently and the other infrequently, then the infrequent feature can be detected through semantic relationship. Recent work shows that semantic relations between terms is most popularly used for identification of concept and features (Cambria *et al.,* 2013; Poria *et al.,* 2013; Weichselbraun *et al.* 2013; Hung & Lin, 2013). Further explanation of the technique is given in the proposed architecture.

2. Related work

As discussed earlier, there are two most frequently reported unsupervised approaches that depend on distribution similarity i.e. association mining (Agrawal & Srikant, 1994) and likelihood ratio test approach (Dunning, 1993). The association mining approach for product features extraction is employed by Hu & Liu (2004). This technique extracts opinion targets through association mining rules. The implementation of this technique was very successful in features extraction. Later on this approach was extended by Wei *et al.* (2010) for the same task with semantic based pruning for refinement of frequent features, identification of infrequent features and for improved the results. The other potentially employed unsupervised classification technique is the LRT. The LRT was introduced by Dunning (1993) and has been reported in different natural language processing (NLP) tasks. The LRT is employed by Yi *et al.* (2003) and Ferreira *et al.* (2008). The LRT technique assumes that a feature related to the topic is explicitly presented by a noun phrase in the document using syntactic patterns associated with subjective adjectives.

Kobayashi *et al.* (2004) used the unsupervised approach for extraction of target features and opinion pairs. This method extracts candidate evaluative expressions using text mining techniques to accelerate the manual annotation process. The authors proposed this method to create an exhaustive list of evaluative pairs for many domains for use as training sets for the machine learning process for feature level opinion mining. Popescu & Etzioni (2005) have introduced web based domain independent system referred as OPINE based on the unsupervised information extraction approach to mine product features from reviews. Carenini *et al.* (2005) developed a model based on user defined knowledge to create taxonomy of product features. Klema & Almonayyes (2006) employed Term Frequency Inverse Document Frequency (TFIDF) and probabilistic classifier to extract fanatic text from blogs. Umagandhi &

Kumar (2014) proposed a supervised heuristic approach for query recommendation. Holziger *et al.* (2006) used domain ontology based on tabular data from web content to bootstrap a knowledge acquisition process for extraction of product features. Zhuang *et al*. (2006) specifically focused on domain of movie reviews for opinion mining. This paper proposed a multi-knowledge based approach which integrates the WordNet, a statistical analysis and a movie knowledge base. Bloom *et al.* (2007) described an unsupervised technique for feature and appraisal extraction. The appraisal expression is a textual unit expressing an evaluative attitude towards some target. Ben-David *et al.* (2007) proposed a Structural Correspondence Learning (SCL) algorithm for domain classification. Lu & Zhai (2008) proposed automatic integration of opinions expressed in a well-written expert review with opinions scattered in various sources such as blogs and forums. This paper proposed a semi-supervised topic model to solve the problem in a principled way. The authors performed experiments on integrating opinions about two quite different topics, i.e. a product and a political review. Kessler *et al.* (2010) presented an annotated corpus containing mentions, co-reference, meronymy, sentiment expressions, and modifiers of sentiment expressions including neutralizers, negators, and intensifiers. Lin & Chao (2010) worked on feature based opinion mining specifically related to hotel reviews. The proposed model used a supervised machine learning approach to train classifiers for tourism-related opinion mining. (Zhai *et al.*, 2011) employed a semi-supervised technique for feature grouping. Goujon (2011) presented a text mining approach based on linguistic knowledge to automatically detect opinion targets in relation to topic elements.

3. Proposed architecture

This section describes the steps of the overall process of the proposed architecture for opinion target identification from unstructured reviews. There are two main objectives for this work; i.e. to improve the candidate selection through dependency patterns and to improve the current LRT technique through semantic relatedness of features, which are employed for opinion target extraction. The architecture explains how opinion targets can be extracted from an input document. This process involves the following three main steps as explained in the block diagram (Figure 1). Each step provides an overview of the sub steps involved in the process.

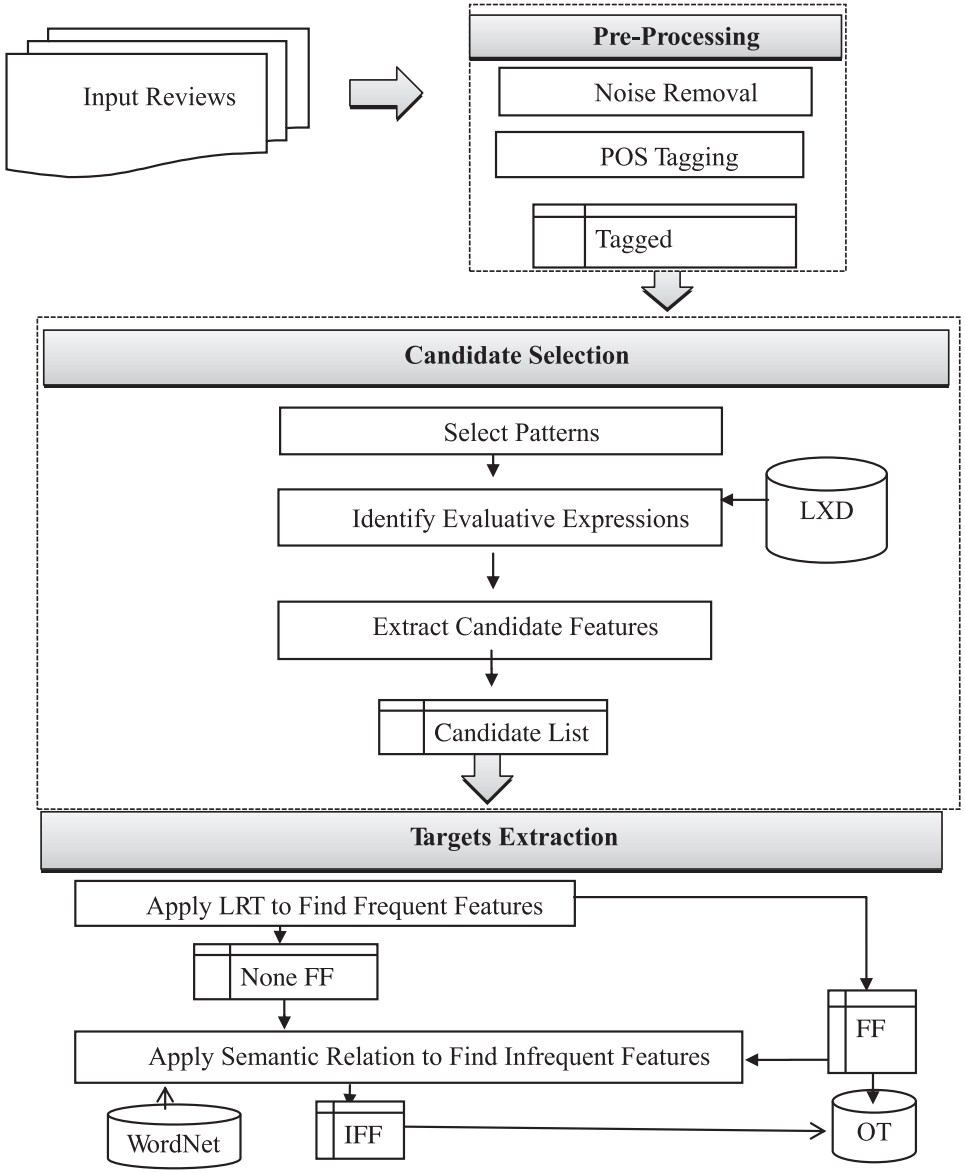

Fig. 1. Proposed architecture for unsupervised learning of opinion targets

3.1. Pre-processing

The first step is to pre-process the input review to make it ready for further processing. This step involves noise removal, part of speech (POS) tagging and sentence splitting. In the POS tagging, each word of the text is assigned a correct grammatical category, which is necessary for pattern generation; e.g. extraction of noun phrases, subjective expressions etc. In this step, noise removal is also performed, which is used to remove incomplete sentences and unidentified words.

3.2. Candidate features selection

This is an important step for opinion target identification, which involves identification of candidate features for opinion target extraction. The proposed algorithm depends on linguistic feature based patterns to identify evaluative expressions containing opinions and targets. In this process, the following two main steps are employed.

3.2.1. Patterns generation

This step involves extraction of strings and expressions based on predefined patterns. The pattern generation is based on the rules explained by Khairullah & Baharum (2012). These patterns are based on base noun phrases with different boundary conditions. The proposed patterns depend on the opinion lexicon dictionary (Hu & Liu, 2004). Hence, the extracted patterns are considered as opinionated expressions, which contain opinion targets.

3.2.2. Candidate selection

In this step, the noun phrases in the extracted evaluative expressions are selected as candidate target features and are ranked on the basis of their frequencies. The candidate opinion targets are further processed to select opinion targets. Hence this is a middle level step to generate a list of candidate opinion targets.

3.3 Targets extraction

The aim of this step is to extract relevant target features from candidate features. In this step, the relevance scoring technique is employed to classify candidate features into relevant and irrelevant. AS described earlier novel SLRT technique is proposed for relevance scoring, which combines the likelihood ratio test (Ferreira *et al.*, 2008) with semantic based similarity scoring. We employed LRT due to its best performance over the other methods as explained earlier. However, the LRT has certain limitations too. For example, as mentioned in the related work, it cannot detect rarely occurred opinion targets. This can be verified from Table 1, which shows a sample of features that have LRT value i.e. (λ=0) due to its low frequency in the given datasets.

Table 1. Sample of rarely occurred opinion targets

Dataset	Features with LRT (Λ=0)
Apex	read, look, sound, price, door, size, design, quality, support, weight, case, forward, output, product, run, unit, video, work, code, direction, disk, display, finish, machine, motor, noise, panel, recognize, service, speed, use
Canon	body, control, depth, design, display, finish, focus, function, image, learning, look, made, noise, option, print, quality, remote, service, shape, shot, speed, use, weight, zoom
Creative	alarm, appearance, balance, break, build, capacity, case, change, clock, control, cover, creative, deal, design, display, equipment, feature, feel, finding, game, look, looking, manage, memory, music, name, option, panel, pause, play, product, program, quality, recognition, recording, remote, remove, style, support, switch, top, unit, use, value, volume, weight, wheel, work, sorting, navigation
Nikon	construction, control, delay, design, function, image, learn, menu, price, quality, size, software, transfer, use, weight
Nokia	application, background, call, command, construction, design, game, keys, look, memory, message, network, picture, plan, quality, resolution, ring, service, software, sound, speaker, tone, use, voice, work

In order to address this issue, semantic based relation between rarely occurred features and frequent features is proposed. Hence the proposed hybrid technique has two steps. In the first step LRT is applied to extract frequently occurred opinion targets, while in the second step semantic based relation is applied to identify rarely occurred opinion targeted. Answer to question; that why semantic based relation is employed to extract infrequent features; it depends on observations and experiments. This technique assumes that the rarely occurred opinion targets have high chance of relation with the frequently occurred features. For example picture, image and photo are closely related; hence if either of the features has occurred frequently and the other infrequently, then the infrequent features can be detected through semantic relationship. Furthermore, a slightly different approach that employs semantic based similarity pruning rule with association mining (Wei *et al.,* 2010) has shown better performance.

In order obtain semantic relation between frequent and rarely occurred feature, the WordNet based IS-A hierarchy is proposed. As explained in the background study, WordNet is a rich lexical source with strong semantic relations between words. WordNet provides different types of relations between words. However the related work proved that, path length similarity in IS-A hierarchy can be potentially employed for entity-to-entity and entity-to-feature relations. Since entity-to-entity or entity-to-feature relations are most likely hierarchical in nature, IS-A hierarchy is ideally

suitable for semantic similarity identification between entities and features. WordNet arranges words hierarchy as shown in Figure 2.

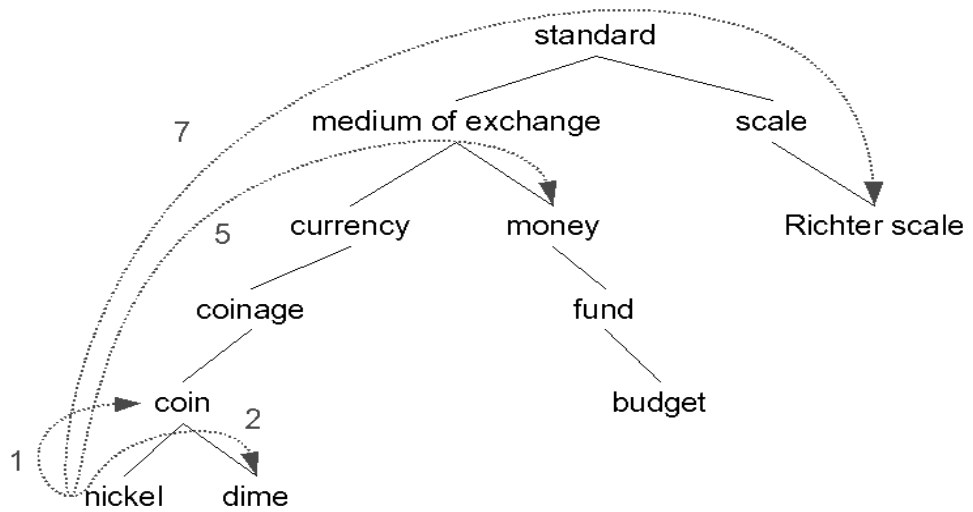

Fig. 2. WordNet IS-A hierarchy (Navigli, 2009)

Path length distance is used to calculate semantic similarity score as follows.

$$Sim(t_1, t_2) = \frac{1}{distance(t_1, t_2)} \quad (1)$$

Where t_1 and t_2 are terms and distance is the path length from t_1 to t_2. It is important to mention why IS–A relations is proposed for semantic similarity in this work. Since this relation do not cross part of speech boundaries, the similarity measures are limited to making judgments between noun pairs (e.g., cat and dog) and verb pairs (e.g., run and walk) (Navigli, 2009). As we only consider noun phrases for opinion targets, IS-A relation is proposed.

The semantic based LRT approach is employed to extract frequent and infrequent features as explained below.

Step 1: In this step, the LRT technique is applied to predict the frequent features. The input to this step is the histogram of the candidate features obtained in step 2. This technique has been formulated by Yi *et al. (*2003) and Ferreira *et al.* (2008) as explained in related work.

Step 2: In the second step, the optimization technique is employed to predict infrequent features based on a semantic relation using the WordNet lexical dictionary. The input to this step is the list of those features, which are classified as irrelevant

by the relevance scoring LRT technique in Step 1. The algorithm in this step finds semantic relatedness of the irrelevant classified features to the relevant features using the WordNet based IS-A relation as explained in Table 2. The IS-A relation is based on the path length similarity between synset (Resnik, 1999) as explained in introduction.

Table 2. Example of WordNet based similarity scores between Frequent and Infrequent features

Document	Frequent Feature	Infrequent Feature	Sim Scores
Nikon Coolpix.txt	Adapters	button	0.88
Nikon Coolpix.txt	Image	photo	0.98
Nikon Coolpix.txt	Flashcards	seconds	0.87
Nikon Coolpix.txt	Camera	camera	0.87

3.4. Tools and implementation

This section provides details about the simulation tools used in this work. For experiments and simulation, the following state-of-the-art software has been employed.

The Stanford part of speech tagger is employed for part of speech tagging (Toutanova & Manning, 2000). This software is freeware and has been widely reported for the part of speech tagging of English language texts. In this thesis, the algorithm is based on the grammatical features of a language element analysis. Therefore, the text of the original dataset are converted to POS tagged corpuses using this software.

TextStat 3.0 is employed for pattern extraction and text analysis. This software is open source and freely available for academic research from the author's website. This software is simple and has been used by a number of works for searching terms and strings in English texts The software accepts any type of regular expression to extract sub strings from a corpus or text documents.

WordNet.Net Library is a set of open source utilities developed by Troy Simpson and is available from the author's website. This library provides a DotNet port to access the WordNet dictionary for similarity scoring. This library is employed in this work for the implementation of the semantic based relevance scoring algorithm. A database of noun phrases with similarity scores is created from all the above mentioned five datasets. From this database, the prototype software extracts the similarity scores to identify the relatedness between frequent and infrequent features.

For the implementation of the proposed semantic based hybrid algorithm, a prototype is developed in VB.Net. The results of the prototype are validated by cross checking

manually with the results of existing approaches. This prototype employs the proposed algorithm to extract features, compares the extracted features with the manually annotated features and creates a confusion matrix based on the comparative results.

4. Results and discussion

4.1. Datasets

This section describes the datasets that have been used for analysis and evaluation in this work. In this work, benchmark datasets of customer reviews about five different products are used. These datasets are crawled from amazon reviews sites and are manually annotated by Hu & Liu (2004). The summary of these datasets is given in Table 3. These datasets are freely available on the author's website and have been reported by a number of research works for product features extraction and opinion summarization. In these datasets each product feature with opinion scoring is properly tagged in each sentence through manual process according to a prescribed annotation scheme as below.

- A sentence is considered as opinionated, if it contains positive or negative comments about features of the product.
- Positive and negative comments are opinion statements containing adjectives that have either positive or negative orientation
- A product feature is the characteristic of the product about which opinions are expressed by the customers.

Table 3. Summary of five products data sets with manually tagged features by (Hu & Liu, 2004)

	Dataset				
	Apex	**Cannon**	**Creative**	**Nikon**	**Nokia**
Reviews	99	45	95	34	41
Total sentences	739	597	1716	346	546
Sentences with target feature(s) and opinion	344(46%)	238(39%)	720(41%)	159(45%)	265(48%)
Total distinct base noun phrases (BNP)	779	811	1641	556	724
Total target features	347	257	736	185	310
Average($\frac{Total\ BNP}{Targets}$)	2.24	3.15	2.22	3.00	2.34
Target types	110	100	180	74	109
($\frac{Target\ types}{Total\ targets}$)	0.32	0.38	0.245	0.40	0.35

4.2. Performance matrices and evaluation criteria

Throughout the experiments the standard performance measures and evaluation criteria have been adopted to ensure the reliability and consistency of the results. The manually annotated datasets have been taken as benchmark for evaluation and comparative analysis.

The three states of the art performance matrices: precision, recall and f-score have been employed for measuring the results of the proposed techniques. To measure these matrices, contingency table of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) counts have been employed as described in Table 4.

The performance measures are calculated as follows.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F - Measure = \frac{2*Precision*Recall}{Precision+Recall} \tag{5}$$

Table 4. Confusion matrix

		Predicted	
		Positive	**Negative**
Actual	Positive	TP	FP
	Negative	FN	TN

4.3. Results

This section explains the results obtained through the proposed algorithm with different experimental setup. The first part is devoted to the results of candidate selection, while the second part examines the results of target feature selection using SLRT with candidate features obtained through different dependency patterns.

4.3.1. Candidate selection

This section presents an extensive analysis of different patterns for candidate selection of opinion target. The purpose of this analysis is to identify the most effective patterns of candidate features for opinion targets. The patterns are mainly based on base noun phrases; therefore, the analysis process is started from base noun phrases. Different combinations of patterns are analyzed in terms of accuracy. The evaluation was

carried out on the benchmark dataset as described above. The output of each pattern is compared with the manually tagged features to identify true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The accuracy rates are evaluated using the confusion matrix. The following subsection presents detailed results of pattern based candidate selection from the reviewed datasets.

- TP=All correctly extracted features
- FP=Extracted BNPs which are not features
- TN=All Non-features BNPs not extracted
- FN=All features BNP not extracted

The results are made clearer through the graph showing the average true positive versus the false positive against each pattern as shown in Figure 3. The average false positive rate of the simple BNP is very high due to no boundary condition with the noun phrase for the candidate selection. The true positive rate of the simple BNP is slightly high, but due to its high ratio of false positives its performance is not reasonable. The false positive rate of the proposed cBNP is very low as compared with the simple BNP; however, its true positive ratio is higher than both the dBNP and bBNP.

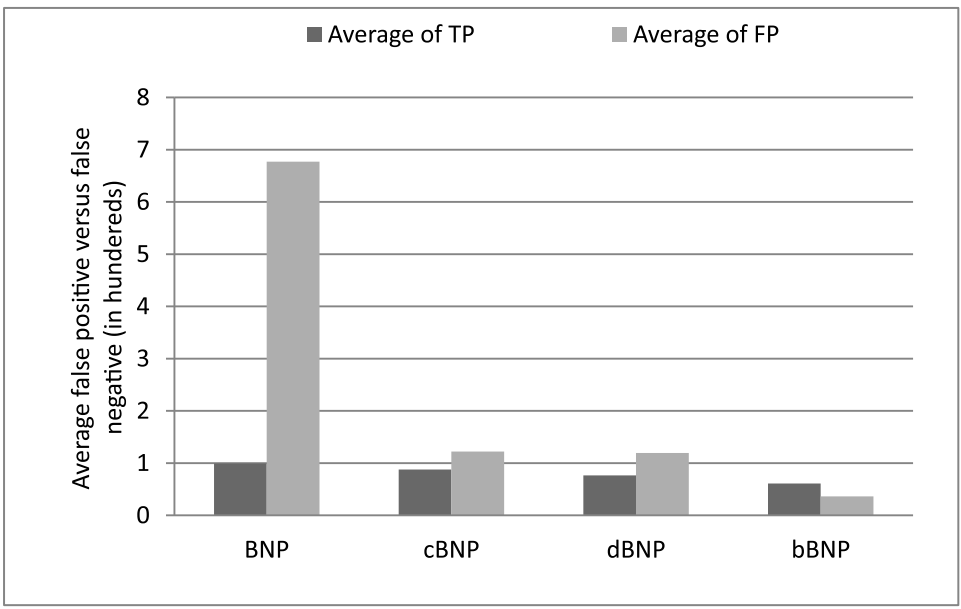

Fig. 3. Pattern based candidate selection: average true positive versus false positive

The overall performance of the four different types of patterns is shown through the graph in Figure 4. The average F-score of the cBNP is high with, balanced precision and recall; hence, the cBNP outperforms the other existing patterns.

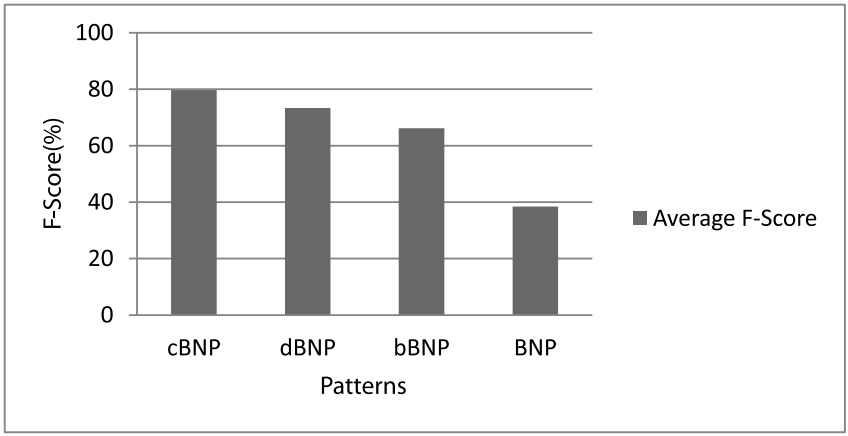

Fig. 4. Pattern based candidate selection: comparison of average F-scores

4.3.2. Opinion target selection

The likelihood ratio test needs two documents for execution. One document is about the topic while the other is a non-topic document. In this setup, the topic documents are the datasets described earlier. The non-topic document is a collection of 600 documents which are collected from the ukWaC, British English web corpus and employed by the existing LRT based opinion target extraction (Ferreira *et al.*, 2008) formulated as follows:

Let D_c denote topic relevant collection of documents and D_n represent collection of documents not relevant to the topic. Then base noun phrases occurring in the D_c are candidate feature to be classified as topic relevant or topic irrelevant using the likelihood ratio test as: if the likelihood score of candidate BNP satisfies the predefined threshold value, then candidate BNP is considered as target feature. The LRT value for any candidate BNP "x" is calculated as:

Let N_1 denote the frequency of a candidate BNP in a Dc, N_2 represent sum of frequencies of all BNPs in D_c except x, N_3 denote frequency of x in D_n, and N_4 represent the sum of frequencies of all BNPs in D_n except the frequency of x. To be more precise, it can be represented with a contingency table as given in Table 5.

Table 5. Contingency table of BNP frequency count

	D_c	D_n
BNP	N1	N2
$\overline{BNP}$	N3	N4

Then the ratios of relevancy of the BNP x to topic and non-topic, which are presented by r_1 and r_2 respectively, can be calculated as below.

$$r_1 = \frac{N_1}{N_1 + N_2} \tag{6}$$

$$r_2 = \frac{N_3}{N_3 + N_4} \tag{7}$$

Thus the combined ratio is calculated as:

$$r = \frac{N_3}{N_1 + N_2 + N_3 + N_4} \tag{8}$$

Hence to normalize the ratios with log:

$$lr = (N_1 + N_2)\log(r) + (N_3 + N_4)\log(1 - r) - N_1 \log(r_1) - N_3 \log(1 - r_1) - N_2 \log(r_2) - N_4 \log(1 - r_2) \tag{9}$$

Hence the likelihood ratio is calculated as below.

$$-2 \log \lambda = \begin{cases} -2 * lr & if\ r_2 < r_1 \\ 0, & if\ r_2 \geq r_1 \end{cases} \tag{10}$$

The likelihood of a candidate BNP is equivalent to the value of λ. Hence the higher value of λ for a candidate BNP has greater chance of relevance and thus considered as target.

The topic documents are converted into POS tagged corpuses using the Stanford parser (Toutanova & Manning, 2000).

We tested the following three different setups with the likelihood ratio test. The definite base noun phrase with likelihood (dBNP-L) and beginning definite base noun phrase with likelihood (bBNP-L) is according to implementation employed by Ferreira *et al.* (2008) while the combined base noun phrase with likelihood (cBNP-L) is based on the proposed cBNP patterns (Khairullah & Baharum 2012).

4.3.2.1. dBNP-L

This setup implements the likelihood ratio test with definite base noun phrase (dBNP) patterns. The dBNP-L method uses candidate features extracted through dBNP and employs the likelihood ratio test for relevance scoring to extract opinion targets.

4.3.2.2. bBNP-L

This setup implements the LRT with beginning definite base noun phrase (bBNP) patterns. The bBNP-L method uses the candidate features extracted through bBNPs and employs the likelihood ratio test for relevance scoring to extract the opinion targets.

4.3.2.3. cBNP-L

This setup implements the LRT with the proposed hybrid patterns (cBNP). The cBNP-L method uses the candidate features extracted through cBNPs and employs the likelihood ratio test for relevance scoring to extract the opinion targets.

To evaluate the performance of each of the above setup, confusion matrix is created as given in Table 6. In this Table, details about the outcomes of the algorithm in terms of true positive, false positive, false negative and true negative are provided. The bold values indicate the best outcomes of the LRT against the patterns with the specified parameters.

Table 6. Details of opinion targets extraction using proposed semantic based LRT

Dataset	Pattern	TP	FP	TN	FN	Total	P(%)	R(%)	F(%)
Apex	dBNP-L	57	19	660	43	779	92.04	51.35	65.92
	bBNP-L	37	8	671	63	779	90.89	33.33	48.78
	cBNP-L	67	16	663	33	779	93.71	60.36	73.43
Canon	dBNP-L	58	330	384	39	811	91.50	51.79	66.14
	bBNP-L	45	13	701	52	811	91.99	40.18	55.93
	cBNP-L	63	42	672	34	811	90.63	56.25	69.42
Creative	dBNP-L	103	69	1415	54	1641	92.51	57.54	70.95
	bBNP-L	78	36	1448	79	1641	92.99	43.58	59.34
	cBNP-L	105	68	1416	52	1641	92.68	58.66	71.85
Nikon	dBNP-L	38	26	466	26	556	90.65	51.35	65.56
	bBNP-L	40	15	477	24	556	92.99	54.05	68.37
	cBNP-L	49	28	464	15	556	92.27	66.22	77.10
Nokia	dBNP-L	62	19	603	40	724	91.85	56.35	69.86
	bBNP-L	56	15	607	46	724	91.58	50.91	65.44
	cBNP-L	72	28	594	30	724	91.99	65.46	74.49

Figure 5 shows the summary results of the LRT based opinion target extraction against each pattern based on the true positive versus false positive. The graph demonstrates that the true positive gain of the cBNP-L is significantly higher than the dBNP-L and bBNP-L, while the false positive of the dBNP-L is higher than the cBNP-L and bBNP-L.

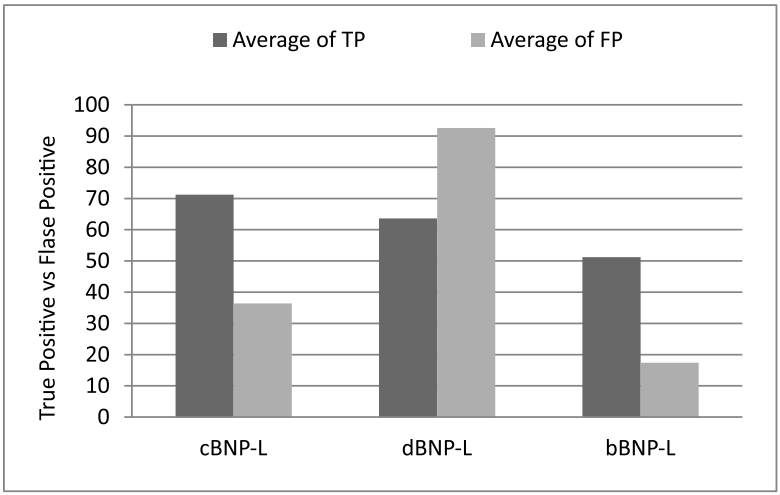

Fig. 5. LRT based features selection: true positive versus false positive

Figure 6 shows the summary results of the LRT based opinion target extraction against each pattern based on true negative and false negative parameters. The graph shows that the true negative rates of the three methods are nearly equal, while the false negative of the bBNP-L is relatively higher than the dBNP-L which is slightly higher than the cBNP-L.

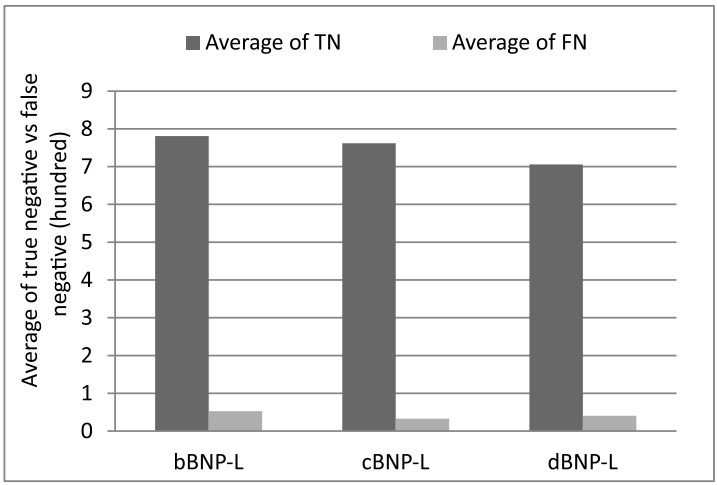

Fig. 6. LRT based features selection: true negative versus false negative

Figure 7 presents the summary results of the LRT based opinion target extraction against each pattern based on precision and recall parameters. The graph demonstrates that the average precision of the three methods are nearly equal, while the recall rate is significantly different. The average recall rate of the bBNP-L is lower than that of the dBNP-L, which is lower than the cBNP-L.

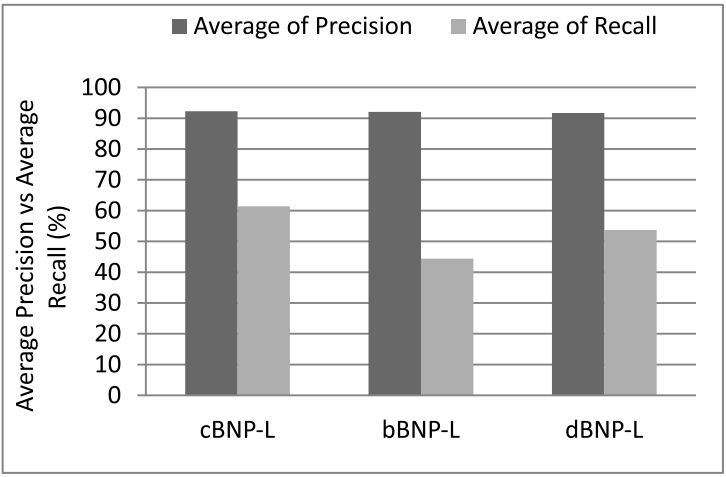

Fig. 7. LRT based features selection: precision versus recall

Table 7 provides the overall summary of the average scores of all the setups based on the precision, recall and f-score. From this table, it is clear that the recall rate of all these setups is not significant other than the best configuration setup; hence, further optimization required to improve the results.

Table 7. Average precision, recall and F-score using SLRT algorithm

Pattern	Measure	Average Score (%)
dBNP-L	Precision	91.71
	Recall	53.68
	F-score	67.69
bBNP-L	Precision	92.09
	Recall	44.41
	F-score	59.57
cBNP-L	Precision	92.25
	Recall	61.39
	F-score	73.26

Figure 8 describes the final outcomes of the likelihood ratio test approach with the three different setups in terms of the f-score. The f-score of the bBNP-L is lower than the dBNP-L, which is lower than the cBNP-L.

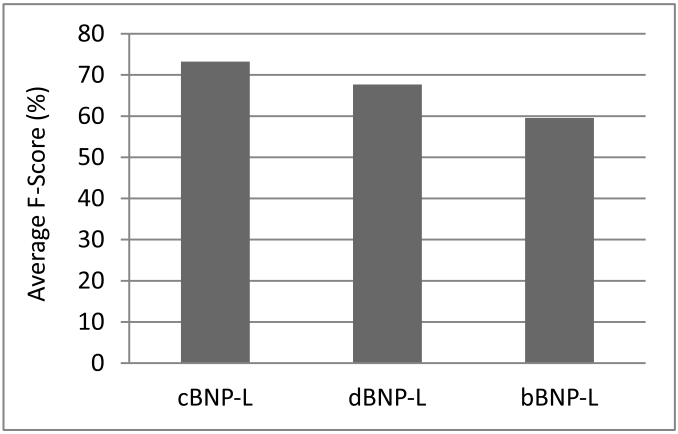

Fig. 8. LRT based features classification: pattern wise average F-score

The first part of the results shows a comparative analysis of pattern and semantic based candidate selection from unstructured reviews. The results of the proposed patterns cBNP based candidate selection are thoroughly compared with the existing pattern based on candidate selection. The Figures 3 and 4 explain the comparative results of candidate selection based on different Lexico-syntactic patterns. The Figure 3 shows that selecting simply all BNPs observe high false positive; on the other hand, restricting the patterns to the article "the" the true positive rate is also decreased with the decrease of false positive. If the lowest false positive rate is considered, then the bBNP performs best; however, its false negative rate is too high and therefore the recall and f-score are comparatively low. The true positive rate of the dBNP is higher than the bBNP and false negative rate is low. Hence, its f-score is higher than the bBNP's. It is observed that the cBNP pattern yields comparatively better results, as its true positive rate is higher and its false negative rate is lower than the other patterns as given in Figure 4.

In the next part of our experiment we tested the target selection through implementation of the lexico-syntactic patterns with three different setups using likelihood ratio test technique. Table 5 shows average precision, recall, and f-score. This table indicates that we have 7.69% improvement in Recall over the existing methods without any loss of precision and hence 5.57% improvements in term of F-Score. The final comparison is explained in graph Figure 8.

4.3.3. Conclusion and future work

This paper presents an in-depth analysis of the pattern features selection for opinion target extraction from unstructured reviews. In the existing literature, it has been found that different associations of base noun phrases are employed for features identification. As all noun phrases cannot be considered as features, certain patterns and rules are used to extract target features. Several patterns have been introduced to restrict noun phrase extraction for features identification. This work presents an analysis of the existing patterns for unsupervised techniques. Analysis shows that certain combination of patterns along with semantic based similarity of frequent and infrequent features outperforms over the existing techniques.

References

Agrawal, R. & Srikant, R. (1994) Fast algorithms for mining association rules in large databases. 20th International Conference on Very Large Data Bases: Morgan Kaufmann Publishers Inc., pp. 487-499.

Ben-David, S., Blitzer J., Crammer. K. & Pereira. F. (2007) Analysis of representations for domain adaptation. In proceedings of Advances in Neural Information Processing Systems, USA, Vol. **137**, pp.1-8

Bloom, K, Garg, N. & Argamon, S. (2007) Extracting appraisal expressions. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics. Rochester, New York, USA, pp 308–315.

Cambria, E., Schuller, B., Xia, Y. & Havasi, C. (2013) New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, (2):15-21.

Carenini, G., Ng, R.T. & Zwart, E. (2005) Extracting knowledge from evaluative text. In Proceedings of the 3rd international conference on Knowledge capture, pp. 11-18, ACM.

Dunning, T. (1993) Accurate methods for the statistics of surprise and coincidence. Comput Linguist, **19**(1):61-74.

Ferreira, L., Jakob, N. & Gurevych, I. (2008) A comparative study of feature extraction algorithms in customer reviews. 2008 IEEE International Conference on Semantic Computing, pp. 144-151.

Goujon, B. (2011) Text mining for opinion target detection. European Intelligence and Security Informatics Conference (EISIC), pp. 322-326.

Holzinger, W, Krüpl, B. & Herzog, M. (2006) Using ontologies for extracting product features from web pages. 5th International Semantic Web Conference, ISWC 2006. Athens, Georgia, USA, pp. 286-299.

Hu, M. & Liu, B. (2004) Mining and summarizing customer reviews. 10th ACM SIGKDD international conference on Knowledge discovery and data mining. Seattle, WA, USA: ACM, pp. 168-177.

Hung, C. & Lin, H.K. (2013) Using objective words in SentiWordNet to improve sentiment classification for word of mouth. IEEE Intelligent Systems, pp. 47-54.

Klema, J. & Almonayyes, A. (2006) Automatic categorization of fanatic texts using random forests. Kuwait journal of science and engineering, **33**(2):1-18.

Kessler, J.S., Eckert, M., Clark, L. & Nicolov, N. (2010) The 2010 icwsm jdpa sentment corpus for the automotive domain. 4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSMDWC2010) Washington, DC, USA.

Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K. & Fukushima, T. (2004) Collecting evaluative expressions for opinion extraction. 1st International Joint Conference on Natural Language Processing. Hainan Island, China, pp. 596-605.

Khairullah Khan. & Baharum, B. Baharudin. (2012) Analysis of syntactic patterns for identification of features from unstructured reviews. 4th International Conference on Intelligent and Advanced Systems (ICIAS), (Volume:1) pp. 165-169.

Khairullah, Khan., Baharum B. Baharudin. & Aurangzeb, Khan. (2013) Mining Opinion Targets from Text Documents: A Review. Journal of Emerging Technologies in Web Intelligence, Vol 5, No 4, pp. 343-353, Nov 2013 doi:10.4304/jetwi.5.4.343-353.

Lin, C.J. & Chao, P.H. (2010) Tourism related opinion detection and tourist attraction target identification. International Journal of Computational Linguistics & Chinese Language Processing, **15**(1):3-16

Lu, Y. & Zhai, C. (2008) Opinion integration through semi-supervised topic modeling. 17th International World Wide Web Conference (WWW '08). Beijing, China. pp. 121-130.

Navigli, R. (2009) Word sense disambiguation: A survey. ACM Comput Survey, **41**(2):1-69.

Popescu, A.M. & Etzioni, O. (2005) Extracting product features and opinions from reviews. Proceedings of the conference on human language technology and empirical methods in natural language processing. Vancouver, British Columbia, Canada: Association for Computational Linguistics, pp. 339-346.

Poria, S., Gelbukh, A., Hussain, A., Das, D. & Bandyopadhyay, S. (2013) Enhanced SenticNet with affective labels for concept-based opinion mining. IEEE Intelligent Systems, pp 31-38.

Resnik, P. (1999) Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, **11**:95-130

Toutanova, K. & Manning, C.D. (2000) Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000). pp. 63-70.

Umagandhi, R. & Kumar, A.S. (2014) Time heuristics ranking approach for recommended queries using search engine query logs. Kuwait Journal of Science, **4**1(2):127-149.

Wei, C.P., Chen, Y.M., Yang, C.S. & Yang, C.C. (2010) Understanding what concerns consumers: a semantic approach to product features extraction from consumer reviews. Info Syst E-Bus Management, (**8**):149-167.

Weichselbraun, A., Gindl, S. & Scharl, A. (2013). Extracting and grounding context-aware sentiment lexicons.IEEE Intelligent Systems, **28**(2):39-46.

Yi, J., Nasukawa, T., Bunescu, R. & Niblack, W. (2003) Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. Third IEEE International Conference on Data Mining (ICDM) pp. 427-434.

Zhai, Z., Liu, B., Xu, H. & Jia, P. (2011) Clustering product features for opinion mining. The fourth ACM international conference on Web search and data mining. Hong Kong, China: ACM. pp. 347-354.

Zhuang, L., Jing, F. & Zhu, X.Y. (2006) Movie review mining and summarization. The 15th ACM international conference on Information and knowledge management. Arlington, Virginia, USA: ACM. pp. 43-50.

Submitted **:** 17/04/2014
Revised **:** 04/01/2015
Accepted **:** 08/01/2015

التحليل الدلالي والنمطي لتحسين تقنيات غير خاضعة للرقابة لتحديد هدف الرأي

خيرالله خان، [1] اشرف الله، [2] باهارومباهارودين [1,*]

[1] معهد الهندسة وعلوم الحوسبة – جامعة العلم والتكنولوجيا – بانو – خيبر باكتونكاوا – باكستان
[2] قسم علوم الحاسوب والمعلومات – جامعة التكنولوجيا – بيتروناس – ماليزيا
[*] مؤلف التواصل : khairullah_k@yahoo.com
بريد المؤلفين: khairullah_k@yahoo.com، ashrafbth@gmail.com

خلاصة

يستخدم التحليل الدلاليو النمطي لتحسين التقنية الحالية لإستخراج الأهداف المتعلقة بالرأي غير الخاضعة للرقابة. ويستخدم خطوتين لتحديد أهداف الرأي: اختيار مرشح وإختيار أهداف الرأي. لاختيار المرشح يتم التعرف على النمط النحوي المعجمي المجمع. أما بالنسبة لأهداف الرأي يستخدم نهجين يجمع بين اختبار نسبة الإحتمال القائمة والشمولية المبنية على دلالات الألفاظ. يقترح النهج الحالي إستخراج الأهداف المتكررة في النص. ومع ذلك، يظهر التحليل أنه لا تحدث كافة ميزات الهدف فيكثير من الأحيان في النصوص.ومن هنا تقترح التقنية الهجينة استخراج كل من الأهداف المتكررة وغير المتكررة. الخوارزمية المقترحة تستخدم نهجا تزايديا لتحسين أداء تقنيات التعدين غير الخاضعة للرقابة عن طريق استخراج ميزات نادرة من خلال القرابة الدلالية مع ميزات متكررة علىأساس قاموس المفردات. تظهر النتائج التجريبية أن تقنية الهجين مع الأنماط الممزوجة تتفوق في الأداء علىالتقنيات الموجودة حاليا.

كلمات مفتاحية: استرجاع المعلومات؛ تعلم الآلة؛ معالجة اللغة الطبيعية؛ تعدين الرأي؛ تحليل النصوص.